Stockholms universitet

Att förnya eller icke förnya

Prediktion av förnyelsegraden inom boendeförsäkringar med logistisk regression och Random Forest

Sanna Kronman

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2017:7
Försäkringsmatematik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Masteruppsats **2017:7**
http://www.math.su.se/matstat

Att förnya eller icke förnya
Prediktion av förnyelsegraden inom boendeförsäkringar med logistisk regression och Random Forest

Sanna Kronman*

Juni 2017

Sammanfattning

Detta masterarbete i försäkringsmatematik undersöker om försäkrings- och kundspecifika egenskaper har en inverkan på om kunden väljer att förnya sin försäkring vid försäkringsperiodens slut. Data vi använder är boendeförsäkringar hos Dina Försäkringar som varit gällande någon gång under 2004 till 2016.

Detta undersöks med två metoder, *logistisk regression* och *Random Forest*. Eftersom andelen ickeförnyelser utgör en sådan liten andel av vårt data tillämpar vi en balanseringsmetod kallad *under sampling*. Denna balansering av data visar sig inte ha en förbättrande effekt i den logistiska regressionsmodellen, men försämrar inte heller resultatet. Balanseringen ger dock en förbättring av predik- tionsförmågan gällande ickeförnyelser i Random Forest-modellen.

Den logistiska regressionsmodellen har något bättre förmåga att prediktera ickeförnyelser än Random Forest-modellen när vi använder allt data från 2004-2015 för att prediktera utfallet 2016, medan Random Forest-modellen är bättre när vi endast använder data från ett år tidigare för att prediktera ett år framåt. Modellerna visar liknande resultat gällande vilka variabler som har en inverkan på förnyelsegraden och inte. Några av variablerna som visar sig ha en inverkan är kundens ålder, försäkringsduration och antal försäkringar kunden har, medan antal skador och medelskadekostnaden inte visar sig ha en stark inverkan.

Eftersom Random Forest-modeller kräver mindre förberedande analyser innan de kan tillämpas och har färre begränsande antaganden gällande data skulle en sådan modell kunna användas som ett analysverktyg för att studera vilka variabler som verkar lämpliga att inkludera i en generaliserad linjär modell, för att på så vis minska tidsåtgången vid modellering.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: sannakronman@gmail.com. Handledare: Mathias Lindholm.

Abstract

This master thesis in insurance mathematics studies whether characteristics concerning the insurance and the customer effects the retention rate within household insurances. We have used data from household insurances that have been active some time in 2004-2016 from any of the insurance companies in Dina Försäkringar.

We have applied a *logistic regression model* and a *Random Forest model* to the insurance data. Since such a small number of our observations were observations of non-renewals we apply a rebalancing method called *under sampling*. This rebalancing did not have any effect on the predictive ability of the logistic regression model, not positive nor negative, while it increased the predictive ability of non-renewals with the Random Forest model.

The predictive ability concerning non-renewals were better with the logistic regression model than the Random Forest model when we use data from 2004-2015 to fit the models and then predict the retention of 2016, while the Random Forest model had a better predictive ability of non-renewals when we used data one year back to predict retention one year ahead. The models gave us similar results of which explanatory variables that had an impact on the retention rate. Some of the variables that had an impact were the age of the customer, the duration of the insurance and how many insurances the customer had, while variables of number of reported claims and average claim cost did not have an impact.

Since the preparation and assumptions regarding data are less when considering a Random Forest model we suggest that such a model could be used as a analytic tool when studying which explanatory variables we should include in a generalized linear model, which then could decrease the time of constructing a model.

Sammanfattning

Detta masterarbete i försäkringsmatematik undersöker om försäkrings- och kundspecifika egenskaper har en inverkan på om kunden väljer att förnya sin försäkring vid försäkringsperiodens slut. Data vi använder är boendeförsäkringar hos Dina Försäkringar som varit gällande någon gång under 2004 till 2016.

Detta undersöks med två metoder, *logistisk regression* och *Random Forest*. Eftersom andelen ickeförnyelser utgör en sådan liten andel av vårt data tillämpar vi en balanseringsmetod kallad *under sampling*. Denna balansering av data visar sig inte ha en förbättrande effekt i den logistiska regressionsmodellen, men försämrar inte heller resultatet. Balanseringen ger dock en förbättring av prediktionsförmågan gällande ickeförnyelser i Random Forest-modellen.

Den logistiska regressionsmodellen har något bättre förmåga att prediktera ickeförnyelser än Random Forest-modellen när vi använder allt data från 2004-2015 för att prediktera utfallet 2016, medan Random Forest-modellen är bättre när vi endast använder data från ett år tidigare för att prediktera ett år framåt. Modellerna visar liknande resultat gällande vilka variabler som har en inverkan på förnyelsegraden och inte. Några av variablerna som visar sig ha en inverkan är kundens ålder, försäkringsduration och antal försäkringar kunden har, medan antal skador och medelskadekostnaden inte visar sig ha en stark inverkan.

Eftersom Random Forest-modeller kräver mindre förberedande analyser innan de kan tillämpas och har färre begränsande antaganden gällande data skulle en sådan modell kunna användas som ett analysverktyg för att studera vilka variabler som verkar lämpliga att inkludera i en generaliserad linjär modell, för att på så vis minska tidsåtgången vid modellering.

Förord och tack

Detta arbete utgör ett examensarbete på masternivå om 30 hp inom Försäkringsmatematik vid matematiska institutionen på Stockholms Universitet.

Jag vill framföra ett tack till Magnus Gustavsson och Fredrik Bendixen som kunde svara på alla mina frågor vid framtagande av försäkringsdata.

Jag vill även tacka Mia Winberg, Anki Koij, Zara Lindberg och Anna Möree som hjälpte till med diskussionen om vilka intervall som var lämpliga att använda i prisförändringen inom försäkringarna.

Tack till aktuarieteamet på Dina Försäkringar AB som varit stöttande och engagerade under min tid som studentaktuarie och som har lett till detta arbete.

Och framförallt ett stort tack till mina två handledare, Mathias Lindholm på Stockholms Universitet samt John Brandel på Dina Försäkringar AB. Tack Mathias för ditt engagemang, inspiration till att studera nya infallsvinklar och din konstruktiva kritik som fått mig att gräva djupare i arbetet. Tack John för idén, diskussionerna och förtroendet till att få skriva mitt masterarbete i samarbete med Dina Försäkringar.

Innehåll

1 Inledning

I detta avsnitt presenteras syftet och metoden som kommer att användas i detta arbete samt bakgrunden till frågeställningen gällande förnyelsegrad.

1.1 Syfte och metod

Syftet med detta arbete är att studera om kund- och försäkringsspecifika faktorer påverkar förnyelsegraden vid försäkringsperiodens slut inom boendeförsäkringarna hos Dina Försäkringar. Vi vill studera detta genom att tillämpa två olika metoder och avgöra vilken som är bäst lämpad för att kunna beskriva om variablerna har en inverkan på förnyelsegraden.

Vi kommer att tillämpa logistisk regression, som följer antagandena gällande generaliserade linjära modeller, och Random Forest, som är en regressionsmetod som bygger på beslutsträd och bootstrap. Dessa två metoder kommer att jämföras i resultat, prediktionsförmåga och lämplighet för att beskriva förnyelsegraden inom boendeförsäkringar.

1.2 Bakgrund

1.2.1 Förnyelsegrad

För alla aktörer som säljer en vara eller en tjänst är det viktigt att förstå sina kunder, för att veta vad de efterfrågar och vad deras behov av varan och tjänsten är. En bättre förståelse för kunden leder till konkurrenskraft på marknaden och möjligheten att utöka sin marknadsandel. På samma sätt är det viktigt för ett försäkringsbolag att förstå sina kunder, för att kunna anpassa sin försäljning och strategi för att nå företagets mål. Den senaste tidens ökade krav på försäkringsbolags rapportering till ansvariga institutioner inom EU samt tydlig information till kunder tros kunna leda till bättre konkurrens på försäkringsmarknaden. Detta ställer högre krav på varje enskilt försäkringsbolags kunskap om marknaden och kunderna.

För att en försäkringsaffär ska växa är det viktigt att utöka beståndet med nya kunder, men det är lika viktigt att behålla de kunder som redan finns i beståndet. Om ett företag har högre kostnader i samband med en ny kund, t.ex. i form av tid för uppläggning av kunden eller information som behöver samlas in, än när en kund väljer att förnya sin försäkring, så kan vinster göras genom att rikta företagets uppmärksamhet mot de befintliga kunderna och hur man ska behålla dem. Det är då viktigt för ett försäkringsbolag att studera förnyelsegraden: andelen som väljer

att förnya sin försäkring efter avslutad försäkringsperiod. Vill vi alltså studera om en kund kommer att förnya eller icke förnya sin försäkring.

1.2.2 Dina Försäkringar

Dina Försäkringar är Sveriges sjätte största försäkringsgivare inom sakförsäkringar (Dina Försäkringar, 2015) som består av 11 stycken sakförsäkringsbolag samt Dina Försäkringar AB. Dina Försäkringar AB ägs av 10 ömsesidiga sakförsäkringsbolag och tillhandahåller dem med service inom IT, försäkringsmatematik, marknadsföring, juridik, m.m. Dina Försäkringar AB äger sedan tillsammans med ägarbolagen Dina Försäkringar Mälardalen AB. Dina Försäkringar AB har även en egen direktaffär av försäkringar som kompletterar lokalbolagen i de geografiska områden där de inte har tillstånd att bedriva egen affär (Dina Försäkringar AB, 2016, s. 4).

Dinabolagen härstammar från brandförsäkringssamarbeten som växte fram i Sverige på 1300-talet. Dessa samarbeten blev sedan flera sockenbolag, där det första startades 1768. År 1987 bildades ett gemensamt återförsäkrings- och servicebolag och 2006 samlades bolagen under det gemensamma namnet Dina Försäkringar (Dina Försäkringar, 2015).

Dina försäkringar erbjuder produkter för privatpersoner, så som person-, motor-, hem- och husdjursförsäkringar, men även försäkringar anpassade för företag och lantbruk. I detta arbete kommer vi att studera boendeförsäkringarna.

2 Teori

I detta avsnitt presenteras den teori som lägger grunden för metoderna som kommer att användas.

2.1 GLM, Generaliserade linjära modeller

GLM är en regressionsmetod som används för att modellera stokastiska variabler. Det är en mycket användbar metod som kan och har tillämpats i de flesta områdena för att få större förståelse om beteende och samband mellan stokastiska variabler.

GLM är, som namnet indikerar, en generalisering av enkel linjär regression. I enkel linjär regression antar vi att den stokastiska variabeln vi vill modellera är normalfördelad. Även om normalfördelningen kanske är den vanligaste förekommande fördelningen inom statistisk teori, så kommer vi många gånger i verkligheten att behöva anta en annan fördelning. Ofta studeras variabler som inte är kontinuerliga, utan antar diskreta värden eller till och med är kategoriska. Om vi vill studera kontinuerliga variabler så har de ibland en skev fördelning istället för symmetrisk. Där blir generaliserade linjära modellen istället användbar då det antas att den stokastiska variabeln vi vill modellera måste tillhöra familjen av exponentialfördelningar. Den omfattar flera fördelningar med många olika egenskaper, däribland normalfördelningen.

2.1.1 De tre komponenterna av GLM

Generaliserade linjära modeller är uppbyggda av tre komponenter: den *stokastiska komponenten*, den *systematiska komponenten* och *länkfunktionen* som alla kommer att beskrivas nedan utifrån s. 116 i Agresti, 2012.

Stokastiska komponenten

Den stokastiska komponenten är den stokastiska variabeln vi vill modellera, också kallad vår *responsvariabel*. Vi kommer att beteckna den som Y och antar att den följer en fördelning inom den exponentiella familjen av fördelningar (se avsnitt 2.1.2). De N stycken observationerna från Y, $(y_1, y_2, ..., y_N)$, antas vara oberoende.

Systematiska komponenten

Den systematiska komponenten är den delen som representerar variablerna som tros kunna förklara responsvariabeln, nämligen våra *förklarande variabler*. Vi har

p förklarande variabler som indexeras av j, $(j = 1, 2, ...p)$ och att x_{ij} utgör observationen från den j:te förklarande variabeln och den i:te observationen. Då kan den systematiska komponenten skrivas som $(\eta_1, \eta_2, ...\eta_N)$ där

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, 2, ...N.$$

Det är parametrarna β_j som behöver skattas i modellen och för att få med ett intercept sätts en $x_{ij} = 1$ för ett j och för alla i och motsvarande β_j brukar då istället betecknas med α.

Länkfunktionen

Den tredje och sista komponenten, länkfunktionen, binder samman de två första komponenterna så att

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, 2, ...N$$

där $E[Y_i] = \mu_i$ och $g(\cdot)$ är en monoton och deriverbar funktion.

2.1.2 Exponentiella fördelningsfamiljen

Exponentiella fördelningsfamiljen (på engelska *exponential family of distributions*) har alla en täthetsfunktion som kan skrivas på formen

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} - c(y_i, \theta_i) \right\}$$

där θ_i är en funktion av väntevärdet, μ_i. Funktionen $a_i(\phi)$ kan oftast förenklas till ϕ/w_i där ϕ är en skalparameter och w_i är vikter tillhörande varje observation (Lindsey, 1997, s. 11-13). Fördelningar som normal-, gamma-, Poisson och Binomialfördelningen är några som tillhör exponentiella fördelningsfamiljen. Vi kan skriva väntevärdet och variansen av vår responsvariabel, Y_i, i termer av $b(\cdot)$ och $a_i(\phi)$ från täthetsfunktionen som

$$\mu_i = E[Y_i] = b'(\theta_i)$$

och

$$Var(Y_i) = a_i(\phi)v(\mu_i)$$

där $v(\mu_i)$ är *variansfunktionen* som definieras som $v(\mu_i) = b''(\theta_i) = b''(b'^{-1}(\mu_i))$ (Ohlsson & Johansson, 2010, s. 23).

2.1.3 Maximum likelihood-skattning

Loglikelihoodfunktionen för en fördelning i den exponentiella familjen blir

$$l(\theta_i; y_i, \Phi) = \log\left(L(\theta_i; y_i, \Phi)\right) = \log\left(\prod_{i=1}^{N} f(y_i; \theta_i, \phi)\right)$$
$$= \sum_{i=1}^{N}\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} - c(y_i, \theta_i)\right\}.$$

För att hitta skattningar på våra β_j-parametrar deriveras loglikelihoodfunktionen med avseende på β_j. Vi kommer då fram till *maximumlikelihoodekvationerna*. Härledningen av dessa ses i Ohlsson & Johansson, 2010, s. 31-32 och blir

$$\sum_{i=1}^{N}\frac{y_i - \mu_i}{a_i(\phi)v(\mu_i)g'(\mu_i)}x_{ij} = 0, \quad j = 1, 2, ..., p.$$

2.1.4 Logistisk regression

Logistisk regression är en typ av GLM där det antas att responsvariabel är binomialfördelad med sannolikheten $\pi(x)$ som okänd parameter och länkfunktionen *logit* används, som definieras av

$$g(\mu_i) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \sum_{j=1}^{p}\beta_j x_{ij}, \quad i = 1, ..., N$$

där $\pi(x)$ då är sannolikheten för att en händelse ska inträffa och måste vara mellan 0 och 1 (Agresti, 2012, s. 119-120). Logit-länkfunktionen kallas också logoddset. Vi antar alltså att $Y \sim Bin(N, \pi(x))$ som har sannolikhetsfördelning

$$p(y) = \binom{N}{y}\pi(x)^y(1-\pi(x))^{N-y}, \quad y = 1, 2, ..., N$$

och väntevärde och varians

$$E[Y] = N\pi(x), \quad Var(Y) = N\pi(x)(1-\pi(x)).$$

Härledningen av maximumlikelihoodekvationerna i en logistisk regression kan ses i Appendix A.

2.1.5 Multikolinjäritet

En effekt som kan uppstå i en statistisk modell är *multikolinjäritet*. Det innebär att det finns ett linjärt samband mellan de förklarande variablerna. Om detta uppstår kan det vara svårt att tolka vilken effekt en förklarande variabel har på en responsvariabel, eftersom den interagerar med en annan förklarande variabel också.

För att studera om det finns multikolinjäritet i en modell kan vi använda oss av ett mått som kallas *GVIF* och står för *Generalized Variance Inflation Factor*. Det definieras som

$$GVIF_i = \frac{\det(\mathbf{R}_{ii}) \cdot \det(\mathbf{R}_{jj})}{\det(\mathbf{R})}$$

där $\det(\cdot)$ står för determinanten och $\mathbf{R}$ är korrelationsmatriser som nu ska beskrivas närmare. Om vi har modellen $\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}$ och låter $\mathbf{X}$ delvis vara en designmatris, så att den både består av kolumner tillhörande kontinuerliga variabler, men att vissa kolumner istället är uppdelade i ettor och nollor för att representera det kategoriska variablerna i modellen. Låt $\mathbf{X}_i$ vara en matris av kolumnerna tillhörande den förklarande variabeln vi vill studera om den har något linjärt samband med resterande förklarande variabler i modellen, så att det kan uppstå multikolinjäritet. De resterande förklarande variablernas kolumner samlar vi i matrisen $\mathbf{X}_j$. Matrisen $\mathbf{R}_{jj}$ är därmed Pearson korrelationsmatrisen för kolumnerna i matrisen $\mathbf{X}_j$, $\mathbf{R}_{ii}$ är Pearson korrelationsmatrisen för $\mathbf{X}_i$ och $\mathbf{R}$ är Pearson korrelationsmatrisen för kolumnerna i hela matrisen $\mathbf{X}$. Pearson korrelationsmatrisen för två vektorer V_m och V_n definieras som

$$\rho_{m,n} = \frac{E[(V_m - \mu_m)(V_n - \mu_n)]}{\sigma_m \sigma_n}$$

där $\rho_{m,n}$ då blir elementet på plats m, n i matrisen, μ_m och μ_n är respektive medelvärde av vektorerna och σ_m och σ_n är respektive standardavvikelse av vektorerna.

För att ta hänsyn till att de förklarande variablerna leder till olika dimensioner av variabler, då vissa är kontinuerliga och andra kategoriska så föreslås att man ska titta på måttet $GVIF_i^{1/2df}$ där df är antalet kategorier i förklarande variabel i minus 1 (Fox & Monette, 1992, s. 180).

Om $df = 1$ reduceras $GVIF_i$ till måttet VIF_i som kan skrivas som

$$VIF_i = \frac{1}{1 - R_i^2}$$

(Fox & Monette, 1992, s. 178) där R_i^2 brukar kallas för förklaringsgraden (på engelska *coefficient of determination*). Ett VIF- eller GVIF-värde större än 5 bör betraktas som att det finns en multikolinjäritet i de förklarande variablerna (Belsley *et al.* , 1980, s. 105).

2.2 Random Forest

Random Forest är en regressions- och klassifikationsmetod vars algoritm grundar sig på *beslutsträd*, *bagging* och *bootstrap*. Vi kommer i detta avsnitt att gå igenom dessa termer och uppbyggnaden av Random Forest för att kunna tillämpa och förstå algoritmen.

Algoritmen Random Forest introducerades av Leo Breiman och Adele Cutler och beskrivs på deras hemsida (Breiman & Cutler, nedladdad 2017-01-25) baserat på rapporten *Random Forest* av Leo Breiman (Breiman, 2001). Nedan förklaring av Random Forests beståndsdelar kommer dock att grunda sig på boken *"An Introduction to Statictial Learning with applications in R"* (James *et al.* , 2013).

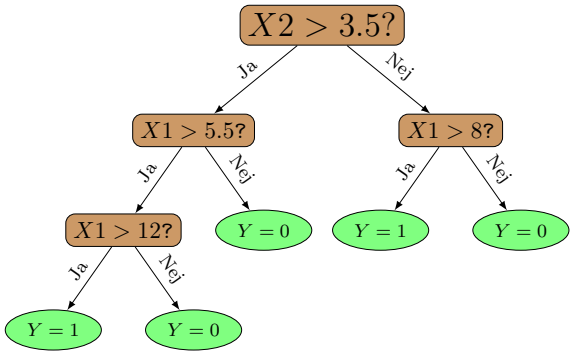
2.2.1 Beslutsträd

För att kunna bygga en skog måste vi börja med ett träd. Precis som i GLM (se avsnitt 2.1.1) så har vi p förklarande variabler, $X = (X_1, X_2, ..., X_p)$ som vi tror kan förklara utfallet på en stokastisk variabel Y i ett beslutsträd. De förklarande variablerna kan vara både kategoriska och numeriska och detsamma gäller för responsvariabeln Y. Om Y är kategorisk kallas beslutsträdet för ett *klassifikationsträd*, medan numeriska värden på Y ger oss ett *regressionsträd*.

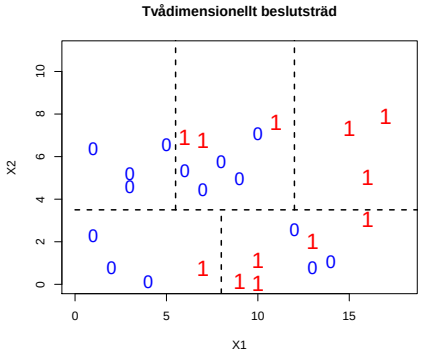
I Figur 1 och 2 ser vi ett exempel på ett tvådimensionellt klassifikationsträd där Y är kategorisk och kan anta värdena 0 och 1 och de förklarande variablerna $X1$ och $X2$ kan anta reella värden mellan 0 och 18 respektive 0 och 15. Givet våra observationer av $X1$ och $X2$ samt ett tillhörande utfall på Y vill vi kunna prediktera ett värde på Y givet nya observationer av $X1$ och $X2$. När vi har observerat $X1$ och $X2$ kan vi i Figur 1 följa beslutsträdet och få en prediktion av Y. I varje brun rektangel ombeds vi att ta ett beslut. Dessa rektanglar utgör det som brukar kallas *grenar* i vårt beslutsträd. Dessa motsvaras av de streckade linjerna i Figur 2. Efter att vi har följt grenarna hamnar vi i en grön ellips, som i ett beslutsträd kallas *löv*. Löven motsvaras i Figur 2 av de slutna områdena som avgränsas av de streckade linjerna (James *et al.* , 2013, s. 305).

När vi når ett löv vill vi veta vad det predikterade värdet på Y är. Detta bestäms olika för om det är ett klassifikationsträd eller regressionsträd. Om Y antar katego-

riska värden så tilldelas Y den kategorin som har majoritet i lövet. Om observerade $X1$ och $X2$ alltså skulle leda oss till området längst ner till höger i Figur 2 skulle det predikteras att Y får utfallet 1. Om Y istället antar numeriska värden så predikteras utfallet av Y som medelvärdet av alla tidigare observationer av Y i lövet (James *et al.* , 2013, s. 311).



Figur 1: *Tvådimensionellt klassifikationsträd, i trädform*



Figur 2: *Tvådimensionellt klassifikationsträd, i grafform*

Ett beslutsträd kan formuleras som modellen

$$f(X) = \sum_{m=1}^{M} c_m \mathbf{1}\{X \in R_m\} = c_i \tag{2.1}$$

där $i = \{j : X \in R_j\}$, eftersom en observation endast kommer att kunna tillhöra ett löv. Det vill säga att $f(X)$ ger det predikterade värdet för kovariatvektorn X baserat på det anpassade trädet, så att $\hat{Y}(X) = \hat{f}(X)$. Vektorn X utgör alltså våra förklarande variabler, M antalet områden, R_m, som finns i trädet, $\mathbf{1}\{.\}$ är en indikatorfunktion och c_m är antingen majoriteten eller medelvärdet i ett löv (James *et al.* , 2013, s. 314).

2.2.2 Att konstruera beslutsträd

För att kunna göra så bra prediktioner med ett beslutsträd som möjligt så finns det metoder för att välja var grenar ska läggas. Placeringen av en gren kommer även att kallas en "split" eftersom en gren delar in ett område i två delar. I Figur 2 har grenarna placerats ut manuellt för att ge ett illustrativt exempel. Om våra förklarande variabler leder oss till området längst upp till vänster ser vi att alla observationer av Y antar värdet 0 och vår prediktion om att Y kommer att anta kategorin 0 är relativt säker. Om våra förklarande variabler istället leder oss till

området längst ner till höger så kommer vi att prediktera kategorin 1 för Y, även om 3 av 8 observationer i området faktiskt tillhör kategorin 0. Det betyder att det oftare i detta område kommer att göra en felklassifikation av Y. Vårt mål vid konstruktionen av ett klassifikationsträd är att göra löven så homogena som möjligt.

Beteckningen $\hat{q}_{mk}$ införs, som anger proportionen av observationerna som är i område m ($m = 1, 2, ..., M$) av kategori k ($k = 0, 1, ..., K$). Om m är området längst ner till höger i Figur 2 så är alltså proportionen av 0-kategori 3/8 och av 1-kategori 5/8. Utifrån $\hat{q}_{mk}$ kan nu måttet *Gini-index* presenteras som

$$G = \sum_{m=1}^{M} \sum_{k=0}^{K} \hat{q}_{mk}(1 - \hat{q}_{mk}).$$

Om ett område, m, är mycket homogent kommer $\hat{q}_{mk}$ i det området att vara nära noll eller ett, vilket leder till att Gini-indexet är litet. Så ett litet Gini-index indikerar homogena löv (James *et al.* , 2013, s. 312). Det finns andra mått som tillämpas för att placera ut grenar i ett klassifikationsträd, t.ex. *cross-entropy*, men i detta arbete kommer vi att använda oss av Gini-indexet.

Om Y antar kontinuerliga eller diskreta värden baseras prediktionen på medelvärdet av observationerna i ett område. Då placeras istället grenar ut så att observationernas avvikelse från medelvärdet i ett område är så liten som möjligt. Vi kan då använda oss av kvadratiska residualsumman (på engelska residual sum of squares, RSS) som ett mått för att avgöra homogeniteten i lövet. RSS ges av

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

där R_j är ett område, $\hat{y}_{R_j}$ är medelvärdet i område R_j och y_i är observation i av Y (James *et al.* , 2013, s. 306).

Utifrån dessa mått på homogenitet eller avvikelser i löven kan vi nu utvärdera olika platser där grenar ska placeras ut. Låt j indexera vilken förklarande variabel som studeras (alltså i vilken "ledd" vi ska lägga ut grenen) och s indexera ett värde på en förklarande variabel (var i den förklarande variabeln, X_j, grenen ska placeras) så kan vi definiera två områden på vardera sida om en gren som

$$R_1(j, s) = \{X | X_j < s\} \text{ och } R_2(j, s) = \{X | X_j \geq s\}.$$

För ett regressionsträd vill vi hitta j och s som minimerar måttet

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

(James *et al.* , 2013, s. 312). Ett motsvarande mått ska minimeras i klassifikationsträdet, där vi använder oss av Gini-indexet.

Efter att en första gren har placerats ut har den delat upp det fullständiga området i två delar. Vi vill därmed minimera måttet igen, men kan nu placera ut en gren som avgränsar det första eller andra området som bildades i och med den första grenen. Så här håller vi på och placerar ut grenar, tills ett lämpligt fördefinierat stoppkriterium nås. Kriteriet kan t.ex. vara att alla områden ska innehålla max fem observationer eller tills vårt RSS inte längre minskar med ytterligare en gren. Eftersom vi väljer ett stoppkriterium själva kan vi konstruera hur stora och komplexa träd som helst. Detta kan dock leda till ett träd som är överanpassat och för komplext. Det finns en metod som kallas *beskärning* (på engelska *pruning*) som reducerar beslutsträdet. Denna metod kommer inte att användas i detta arbete och den intresserade läsaren hänvisas därför till s. 307 i *An Introduction to Statistical Learning with Applications in R* (James *et al.* , 2013).

2.2.3 Bagging

Bagging grundar sig på idén med bootstrap. I bootstrap simuleras nya dataset från de ursprungliga observationerna för att på så vis få många dataset att t.ex. testa modeller med eller för att beräkna variationen i skattade parametrar (James *et al.* , 2013, s. 187-189).

Ett beslutsträd kan vara en väldigt instabil regressionsmetod. Om vi får ytterligare en observation så kan konstruktionen av trädet ändras mycket. Eller om observationerna skulle dela upp i två dataset och ett beslutsträd skulle anpassas till vardera dataset kan de två beslutsträden se väldigt olika ut. Vi kan använda oss av bagging för att reducera denna instabilitet.

Om vi har n oberoende stokastiska variabler, $(Z_1, Z_2, ..., Z_n)$, som alla enskilt har variansen σ^2 så kommer medelvärdet av dessa variabler, $\bar{Z}$, att ha variansen σ^2/n. Variansen kommer alltså att minska med antalet variabler. Därför konstrueras flera beslutsträd, $(\hat{f}^1(x), \hat{f}^2(x), ..., \hat{f}^B(x))$ (se ekvation 2.1), baserat på B olika dataset av observationer, alla slumpade från det ursprungliga och fullständiga datasetet i enlighet med bootstrap. Sedan tas medelvärdet av prediktionerna från de B beslutsträden för att göra en prediktion med lägre varians,

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x) = \frac{1}{B}\sum_{b=1}^{B}c^b$$

där $c^b = \{c_j^b : x \in R_m^b\}$ och b alltså är ett index för vilket beslutsträd det är och c_j^b kommer från ekvation 2.1 (James *et al.* , 2013, s. 316). Ovan medelvärde går att beräkna när prediktionerna av Y är diskreta eller kontinuerliga. Om Y istället antar kategoriska värden så registreras alla prediktioner som har gjorts av Y i de B anpassade klassifikationsträden och väljer en slutgiltig prediktion av Y som majoriteten bland prediktionerna från de B träden. Det blir en majoritetsröstning i två steg: majoriteten i varje enskilt löv i träden, sen majoriteten bland träden (James *et al.* , 2013, s. 317).

2.2.4 Val av förklarande variabler i ett beslutsträd

En nackdel som uppstår med bagging är att vi förlorar en del av de tolkningar vi kan göra genom att konstruera *ett* beslutsträd. T.ex. kan vi i Figur 1 och 2 se att högre värden på $X1$ och $X2$ ser ut att ge upphov till kategorin 1 på responsvariabeln Y, medan låga värden ger upphov till kategorin 0. Detta är vår tolkning av ett träd, men samma tolkning kommer inte kunna göras för alla våra konstruerade träd i bagging. Dock kan en del av den informationen sammanfattas genom att konstruera många träd i och med bagging (James *et al.* , 2013, s. 319).

I Avsnitt 2.2.2 introduceras RSS och Gini-index som används och minimeras när beslutsträd ska konstrueras. För varje gren som placeras ut kan vi registrera vilken förklarande variabel som grenen placerades vid och hur mycket just den grenen minskade RSS och Gini-index. På så vis kan vi få en bild av vilka förklarande variabler som är viktiga att dela in i flera områden för att få en bra prediktion av Y (James *et al.* , 2013, s. 319). Om det t.ex. aldrig placeras en gren så att den förklarande variabeln X_j är den som ger upphov till att Y får en specifik prediktion kan det antas att X_j faktiskt inte har en så står inverkan på den stokastiska variabeln Y.

2.2.5 Out-of-bag felskattning

När vi i enlighet med bootstrap slumpar ett dataset från våra ursprungliga observationer med återläggning så kommer inte alla observationer att ha använts för att konstruera beslutsträdet. De observationer som inte har använts för det trädet kallas *out-of-bag observationer* (OOB-observationer). Dessa kan användas för att utvärdera prediktionerna i beslutsträden.

För varje observation i används de beslutsträd där observation i är en OOB-observation för att prediktera ett tillhörande värde på Y. Vi får lika många prediktioner för observation i som antal träd där den observationen var en OOB-observation och dessa prediktioner kan vi ta medelvärdet av eller majoriteten för att få en prediktion som kallas OOB-prediktionen. En sådan prediktion kan fås för alla n observationer i vårt ursprungliga dataset och baserat på dessa kan det beräknas ett kvadratisk medelfel (på engelska mean squared error, MSE), för beslutsträdet konstruerat med bagging. MSE definieras som

$$MSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

där y_i är det faktiska observerade utfallet och $\hat{y}_i$ är OOB-prediktionen (James *et al.* , 2013, s. 317-318).

2.2.6 Algoritmen Random Forest

Med hjälp av teorin från Avsnitt 2.2.2 och 2.2.3 kan nu algoritmen för Random Forest konstrueras.

Observationer slumpas från vårt ursprungliga data med återläggning, så att vi får nya dataset. Baserat på dessa dataset konstrueras ett beslutsträd för vardera set. Vid konstruktionen av beslutsträden lägger vi dock in ytterligare en dimension av bootstrap, nämligen vid val av förklarande variabler som grenarna ska placeras vid. Om vi egentligen har möjlighet att välja bland p förklarande variabler att placera en gren vid så slumpas m av dessa ut som kandidater för en split av ett område. För varje ny gren som ska placeras ut så slumpas igen m nya kandidater ut från hela setet av förklarande variabler. Ett vanligt val av antalet m är $\sqrt{p} = m$. Observera att om $m = p$, alltså att det kan väljas bland alla förklarande variabler i varje split, så blir Random Forest-algoritmen bara "ren" bagging igen (James *et al.* , 2013, s. 320).

Genom att göra detta kommer vi att få en lägre korrelation mellan prediktionerna som görs eftersom trädens grenar tvingas att gå olika vägar. Tänk er t.ex. att om en förklarande variabel har en stor inverkan på prediktionen av Y så kommer alla träd som konstrueras att snabbt placera ut en gren eller flera som delar in denna variabel, eftersom den indelningen minskar RSS eller Gini-index mycket. Det leder in konstruktionen av trädet på samma väg för alla träd och gör dem mer lika. För att variationen för prediktionen (som är ett medelvärde eller majoritet av prediktioner) ska minska, så som förklarat i Avsnitt 2.2.3, så vill vi att prediktionerna från alla träd ska vara så oberoende som möjligt.

Algoritmen kan sammanfattas i följande steg:

Algoritmen Random Forest

- -

Steg 1. Slumpa ett bestämt antal observationer från vårt ursprungliga data med återläggning.

Steg 2. Konstruera ett beslutsträd baserat på observationerna från Steg 1 genom att vid varje gren som ska placeras välja ut m av de p förklarande variablerna som är möjliga att placera en gren vid.

Steg 3. Stanna konstruktionen av trädet när lämpligt stoppkriterium har nåtts, t.ex. antal observationer i löven eller antal grenar.

Steg 4. Repetera Steg 1-3 för så många träd vi vill ha.

- -

2.3 Utvärdering av modellprediktion

En modells prediktiva förmåga kan beskrivas med hjälp av en ROC-kurva (*Receiver Operating Characteristic Curve*). När en modell har byggts så kan den testas genom att använda observationer där vi redan vet utfallet av responsvariabeln för att se om modellen lyckas prediktera samma utfall. Modellerna kommer att prediktera en sannolikhet för ett utfall, t.ex. $P(Y = 1|X)$. För att modellen ska göra en prediktion bestäms en tröskel, c, så att prediktionen att $Y = 1$ görs om till $P(Y = 1|X) > c$. I en modell med binära utfall så kan prediktionen modellen gör sammanfattas i Tabell 1, en klassifikationstabell.

		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	*TP*	*FP*
	$Y = 0$	*FN*	*TN*
	Total	*P*	*N*

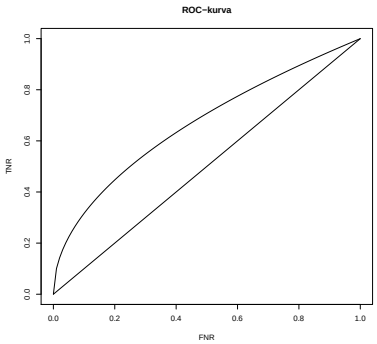
Tabell 1: *Möjliga prediktionsutfall. TP = True Positive, FP = False Positive, FN = False Negative, TN = True Negative.*

Antalet prediktioner som testas är $N_p = TP + FP + FN + TN$. Tabellen tolkas så att av alla observationer som faktiskt har utfallet 1 så predikterades TP av dem

som 1 och av de som faktiskt hade utfallet 0 så predikterades FP av dem som 1. Den allmänna *klassifikationsgraden*, CR, i modellen blir $(TP + TN)/N_p$. Detta mått brukar även kallas *Accuracy*. Vi definierar även måtten, *True Negative Rate* samt *False Negative Rate* som

$$TNR = \frac{TN}{N} \quad och \quad FNR = \frac{FN}{P}.$$

Eftersom prediktionen beror på tröskelvärdet c så kommer dessa mått att variera med c. Därför kan TNR ritas upp mot FNR i en graf för alla värden på c som alltså går mellan 0 och 1. Denna graf kallas ROC-kurva och ytan under denna kurva beskriver modellens förmåga att skilja på de som får utfall $Y = 1$ mot de som inte får det. Ytans storlek går mellan 0.5 och 1 och ju större yta desto bättre förmåga att skilja mellan utfallen (Hosmer *et al.* , 2013, s. 169-177). Ett exempel på en ROC-kurva ses i Figur 3.



Figur 3: *Exempel på en ROC-kurva. Den diagonala linjen är för att illustrera det lägsta värdet på ytan under en ROC-kurva.*

3 Data

I detta avsnitt presenteras vilket data som kommer att användas. Vi går igenom vad observationerna innebär, vilka förklarande variabler vi har, antaganden och begränsningar som används.

3.1 Försäkringsdata

Data som kommer att användas i det här arbetet är boendeförsäkringar tecknade hos någon av de 11 lokalbolagen inom Dina-federationen eller Dina Försäkringar AB. Försäkringarna har varit gällande någon gång under tiden 1a januari 2004 till och med 31 februari 2017. En observation utgör en försäkringsperiod för en boendeförsäkring, som normalt varar ett år om den inte annulleras. Vi tar inte med försäkringar som är mitt inne i en försäkringsperiod, utan endast historiskt data över avslutade försäkringsperioder. Om jag som kund till exempel köper en boendeförsäkring den 1a oktober 2013 och även väljer att förnya den två år i rad så kommer jag att ha haft försäkringen i tre försäkringsperioder och totalt haft försäkringen i tre år. Den första observationen är försäkringsperioden som löper från den 1 oktober 2013 till 31 september 2014. Den andra från 1 oktober 2014 till 31 september, osv.

Varje observation har information om försäkringen och kunden i olika variabler som kommer att presenteras i Avsnitt 3.2. Totalt har vi 1 247 327 antal observationer.

Se avsnitt 3.3 för vilka begränsningar som har gjorts i försäkringsdata samt avsnitt 4.4 och 4.3 för diskussion kring vilka observationer som kommer att användas för modellering i olika modeller.

3.2 Variabler

Vi ska titta på om egenskaper hos en kund och dess försäkring kan tänkas påverka sannolikheten att kunden väljer att förnya sin försäkring efter en försäkringsperiod. Vad skulle kunna tänkas påverka viljan att fortsätta tillhandahålla ett försäkringsskydd hos ett specifikt bolag?

Bilden av en trogen kund är en kund som har många försäkringar inom samma bolag och som har varit kund länge. Därför studeras variabler som just beskriver hur länge kunden har haft en försäkring hos ett Dina-bolag och hur många försäkringar kunden har. Även försäkringar utöver den produkttyp vi har valt att studera räknas. Om en kund innehar en boendeförsäkring och en motorförsäkring

samma år så räknas det som att kunden innehar två försäkringar. Vi tittar även på mer specifika egenskaper om kunden så som kön, ålder och om kunden bor i storstad, tätort eller landsbygd, men också via vilken försäljningskanal kunden har köpt sin försäkring och hur kunden väljer att betala sin försäkring.

En intressant variabel att studera är prisförändringen. Ett antagande som känns rimligt är att ju dyrare en försäkring blir, desto mindre benägna är vi att köpa den. Måttet som avgör hur pass mycket färre försäkringar vi köper när priset på dem stiger kallas *priselasticitet*. Därför tas prisändringen med för att studera om förändringen i priset på försäkringen har en inverkan på sannolikheten att kunden väljer att förnya sin försäkring.

Vi vill även studera om information om skadorna som är kopplade till försäkringarna har en inverkan på förnyelsegraden. Variabler om totala antalet skador och medelskadekostnaden som har uppkommit under den fullständiga tiden som försäkringen har varit gällande samt antalet skador och medelskadekostnaden under den nuvarande försäkringsperioden tas med. Detta görs för att studera om en kund som har varit i kontakt med ett bolag i och med en skada visar större sannolikhet att förnya, än en kund som inte har haft någon kontakt med bolaget.

En variabel tas med som berättar vilket system som har använts för att hantera försäkringsinformationen. Dina Försäkringar AB har under tiden för de observationerna vi har bytt system för hur försäkringarna hanteras. Därför tas en variabel med för att kontrollera så att vi inte ser för stora skillnader i data som kan vara orsakat av systemövergången.

Många variabler har tagits med som eventuellt kan beskriva förnyelsegraden och att använda dem alla kan leda till en väldigt komplex modell. Anledningen till att vi har valt att studera så många är för att se hur GLM och Random Forest väljer ut vilka variabler som har en signifikant inverkan på förnyelsegraden och för att jämföra resultaten. På så vis utmanas metoderna med många variabler.

Variablerna kommer att pseudonymiseras med namnen `variabel1`, `variabel2`, osv. och det kommer inte att föras en diskussion gällande variablernas egenskaper och dess utfall, på grund av sekretesskäl.

3.3 Begränsningar och antaganden i försäkringsdata

En variabel som ska undersökas är om prisförändringen av en försäkringspremie har någon inverkan när kunden väljer att förnya sin försäkring eller inte. Vi vill därför särskilt fånga upp de kunder som har fått förfrågan om förnyelse och blivit

presenterade en premie för sin kommande försäkringsperiod och ifall de då har valt att förnya eller inte. Förfrågan om förnyelse av försäkringen och kommande premie presenteras för kunden 40 dagar innan förnyelsedatum. Om en ny försäkringsperiod har startats på försäkringen är det klart att kunden har accepterat premien och valt att förnya. Om kunden aktivt har avslutat sin försäkring eller inte har betalat sin premie inom en månad efter förnyelsen väljer vi att tolka det som att kunden har tagit del av premien för den nya försäkringsperioden och därefter valt att inte förnya sin försäkring. De som har annullerat sin försäkring har exkluderats ur data, vilket diskuteras närmare i Avsnitt 4.4.

Observationerna består av kontrakt som studeras på "försäkringsnivå", medan försäkringen i sin tur består av olika objekt och risker. T.ex. så kan en boendeförsäkring innehålla en villa och ett fritidshus. Om en kund med endast en villa under året köper ett fritidshus och vill utöka sin försäkring i samband med förnyelsen så kommer premien att öka betydligt från den tidigare försäkringsperioden. Vi kommer vid förnyelsen att se denna stora ökning av premien och även se att kunden har valt att förnya. Kunden blir här eventuellt inte avskräckt av ökningen, eftersom kunden vet att flera objekt nu är försäkrade. Därför utesluts försäkringar där antalet objekt har förändrats från innan förnyelsen och efter, och även om antalet objekt har förändrats från försäkringens start till nuvarande försäkringsperiod, eftersom det mycket troligt innebär en förändring av premien som kunden också kan anse motiverad. Eftersom försäkringen innehåller mer än tidigare så accepterar kunden därför en prisökning. Ett alternativ hade varit att studera försäkringarna på "objektsnivå", men på grund av systembytet som gjorts så finns inte möjligheten att följa objekten under en längre tid på samma sätt som vi kan följa en försäkring under en längre tid.

Flera förändringar i försäkringen kan göras som leder till ökningar eller minskningar i premien, som kunden också skulle kunna anse motiverade just på grund av förändringen som meddelas. T.ex. kanske delar har byggts till av ett hus och därmed förändrat premien. Eftersom sådana här förändringar är svåra att följa väljer vi endast ut försäkringar där prisförändringar befinner sig inom en visst intervall och då troligare har uppkommit av premieförändringar på grund av försäkringsbolagets tariffer än av aktiva förändringar från kundens sida. Intervallet som kommer att användas är en premieförändring inom -10% och +15% vid en ny försäkringsperiod. Det kommer även att göras ett begränsningsintervall baserat på hur mycket premien har förändrats från första försäkringsperioden till nuvarande försäkringsperiod. Om premien från start har minskat med mer än -50% eller ökat med mer än 100% så utesluter vi observationen. Intervallet är något större uppåt för att det oftare sker en ökning av premien än en minskning.

Det finns information om kunder och försäkringar från år 2001, men på grund av vissa avvikelser i data de första åren så väljer vi att endast studera försäkringar som har sin försäkringsperiod från och med 1 januari 2004. Vi har dock information om hur länge kunden har varit försäkrad inom ett Dina-bolag samt hur länge försäkringen har funnits ända från 1a januari 2001.

Utöver detta så har observationer där vi saknar information uteslutits och vi studerar endast kunder med åldrar mellan 16 och 90 år. Vi exkluderar även kunder som har haft fler än 20 försäkringar samma år eftersom en sådan kund inte representerar en genomsnittlig försäkringskund.

4 Modellering

I detta avsnitt kommer vi att resonera kring uppbyggnaden av de statistiska modellerna som ska användas och analyseras samt metoderna som kommer att användas. Metoderna bygger på teori som har presenterats i avsnitt 2.

4.1 Är en försäkring priselastisk?

I idéstadiet av detta arbete studerades möjligheten att modellera *priselasticiteten* hos försäkringskunder. Priselasticiteten definieras som

$$\mathcal{E}_{t,i} = -\frac{\frac{Q_{t,i}-Q_{t-1,i}}{Q_{t-1,i}}}{\frac{P_{t,i}-P_{t-1,i}}{P_{t-1,i}}} \tag{4.1}$$

och är alltså den procentuella förändringen i efterfrågad kvantitet (Q) dividerat med procentuell förändring i pris (P), multiplicerat med -1. Index i utgör t.ex. en särskild grupp försäkringar eller en specifik försäkring, medan t är ett tidsindex. Priselasticiteten är oftast positiv: ju större ökning i priset desto mindre kvantitet av varan efterfrågas. Om priselasticiteten är hög så kallar man varan priselastisk, medan en låg priselasticitet gör en vara inelastisk.

En första fundering är då om boendeförsäkringar är priselastiska varor? Bryr sig försäkringskunder om priset på sina boendeförsäkring tillräckligt mycket för att avsluta den vid en prisförändring? En genomsnittlig försäkringskund skulle troligtvis inte på rak arm kunna svara på vad deras försäkring kostar dem per år, så att en kund skulle välja att avsluta sin försäkring vid en relativt stor prisökning känns inte heller självklart. Detta skulle till stor del kunna bero på att försäkringar är "sällan köps-varor" och att vi oftast bara kommer i kontakt med priset för vår försäkring en gång om året, vid förnyelse av försäkringen, och då inte har koll på marknadspriset i övrigt. Detta till skillnad från t.ex. en matvara som vi handlar varje vecka och då oftare kommer i kontakt med priset på varan. Att då konstruera en modell som ska avgöra om olika egenskaper och faktorer hos en försäkringskund påverkar kundens priselasticitet känns svårmotiverad, om vi inte först har besvarat frågan: Är en boendeförsäkring priselastisk?

Ytterligare en aspekt som gör priselasticiteten av försäkringar svår att mäta är att priset många gånger är konstant för en försäkring över tiden. Om vi alltså studerar en försäkring för en försäkringskund, där priset vid förnyelse inte har förändrats, så blir nämnaren 0 i ekvation 4.1 och vi kan inte använda den observationen. Dock skulle en eventuell aggregering av data eller transformation av måttet priselasticitet kunna göras för att kringgå detta problem och därmed kunna studera

priselasticiteten som responsvariabel.

Om ekvation 4.1 närmare studeras och vi frågar oss vad som är stokastiskt så ses att det endast är termen $Q_{t,i}$ som är stokastisk. Vi vet om en kund har haft en försäkring tidigare, eller hur många försäkringar det finns i en viss grupp ($Q_{t-1,i}$), och vi vet den procentuella förändringen av priset för den enskilda försäkringen eller genomsnittligt i en grupp av försäkringar (nämnaren i ekvation 4.1). Det känns därmed motiverat att fokusera på modellering av kvantiteten $Q_{t,i}$ istället. Egenskaper hos $Q_{t,i}$ och hur denna kvantitet är relaterad till förnyelsegraden presenteras i avsnitt 4.2. Vi kommer därför att bygga upp en modell som studerar om egenskaper och faktorer hos en försäkring och försäkringstagare påverkar förnyelsegraden istället, alltså andelen som väljer att förnya sin försäkring. En av dessa faktorer som kommer att inkluderas i modellen är prisförändringen. På så vis kan vi få svar på om prisförändringen har en inverkan på om en kund väljer att förnya sin försäkring eller inte och ge en indikation på om en boendeförsäkring är priselastisk.

4.2 Förnyelsegrad

Förnyelsegraden definieras, inom försäkringsbranschen, som andelen kontrakt av de som påbörjade en försäkringsperiod som finns kvar vid nästa påbörjade försäkringsperiod. Om vi låter $Q_{i,t}$ vara antalet kontrakt i grupp i vid tiden t så blir förnyelsegraden

$$g_{i,t} = \frac{Q_{i,t}}{Q_{i,t-1}}.$$

Indexet i kan utgöra en viss försäkringsprodukt, ett visst bolag eller gå ner enda på försäkringskontraktsnivå. Om vi studerar ett enskilt försäkringskontrakt så kommer $Q_{i,t}$ endast att kunna anta värdet 0 (om försäkringskontraktet inte förnyas) eller 1 (om försäkringskontraktet förnyas). $Q_{i,t-1}$ är då alltid 1. Om vi istället tittar på alla kontrakt inom en försäkringsprodukt så kan $g_{i,t}$ vara alla reella tal mellan 0 och 1. Vid 0 så har alla valt att inte förnya sitt försäkringskontrakt och vid 1 så har alla valt att förnya sina kontrakt.

I variabeln $g_{i,t}$ är det täljaren som är stokastisk. Vi vet inte hur många som kommer att vilja förnya sitt kontrakt. $Q_{i,t}$ kommer att vara binomialfördelad, eftersom vi undrar hur många av $Q_{i,t-1}$ antal kontrakt som kommer att förnyas. Sannolikheten för förnyelse är $\pi_{i,t}$ och därmed kan det skrivas att $Q_{i,t} \sim Bin(Q_{i,t-1}, \pi_{i,t})$. Den okända parametern är $\pi_{i,t}$ och eftersom denna parameter säger vad sannolikheten för förnyelse är så kommer detta också att vara skattningen på vår förnyelsegrad.

4.3 Oberoende observationer

När man skapar en modell med GLM så är det viktigt att observationerna som modellen skattas på är oberoende (se asvnitt 2.1.1). Detta är för att vi endast vill modellera hur de förklarande variablerna i modellen påverkar utfallet av responsvariabeln och inte att denna effekt "skuggas" av hur en annan observation påverkar utfallet. Därför skulle det vara problematiskt att studera två olika försäkringar med samma försäkringstagare i en modell. En hypotes är att om en kund t.ex. innehar två försäkringar och efter en försäkringsperiod väljer att avsluta en av försäkringarna, så ökar sannolikheten för att kunden även avslutar sin andra försäkring. Vi undviker detta eventuella samband genom att endast studera en sorts försäkring per försäkringskund, i det här fallet boendeförsäkringar.

Ytterligare en fråga om oberoende i data uppstår när vi, som i det försäkringsdata vi har, har flera försäkringsperioder för en och samma försäkring. En försäkring som har varit gällande under hela 2004 och 2005 har då varit gällande två försäkringsperioder och om kunden inte under 2004 hade valt att förnya sin försäkring så hade inte ens observationen av försäkringen från 2005 existerat. Detta kallas en *longitudinell studie* där man har följt ett objekt under en längre tid och vid flera tillfällen samlat in information om det objektet, i det här fallet ett försäkringskontrakt och försäkringstagaren för det kontraktet. Ett problem som då kan uppstå är om en specifik kund har en större tendens till att förnya eller inte. Vi skulle alltså kunna anta att det finns en kundspecifik effekt, eller kanske snarare kontraktsspecifik effekt, som då skulle påverka alla observationerna för det kontraktet.

En metod för att hantera en sådan effekt skulle vara att införa en *slumpmässig effekt*-variabel, som är specifik för varje kontrakt. En sådan modell kallas på engelska en *mixed model* eller *random effects model*. Om många kontrakt skulle ha väldigt starka specifika effekter som man bör ta hänsyn till i en statistisk modell, men inte gör det, kan det leda till att kovariansmatrisen för de skattade parametrarna är inkorrekt. Det leder i sin tur till att konfidensintervall och test gällande parametrarna blir fel (Fahrmeir *et al.* , 2013, s. 349-353). Vi väljer i det här arbetet att anta att det inte finns några starka kontraktsspecifika effekter, eller att effekter från olika kontrakt tar ut varandra. Alltså kommer data bestående av flera observationer från samma kontrakt att användas och vi får i utvärderingen av den generaliserade modellen ha denna eventuella kontraktseffekt i åtanke.

En annan metod för att modellera förnyelsegraden och kunna ta tillvara på alla våra observationer utan att riskera ett problem med beroende bland observationerna skulle vara att tillämpa överlevnadsteori. Vi skulle då kunna se flera av kovariaterna som beroende av tid och använt försäkringsduration som exponeringstiden,

men denna metod kommer inte att tillämpas i detta arbete.

4.4 Annullation eller ickeförnyelse

När en kund väljer att avsluta sin försäkring sker detta antingen under försäkringsperioden eller vid slutet av försäkringsperioden när kunden har blivit erbjuden en förnyelse. Vi kan skilja på dessa två händelser som *annullation* (sker under försäkringsperioden) samt *ickeförnyelse* (sker vid erbjuden förnyelse). Vid ickeförnyelse så har kunden blivit erbjuden en premie för kommande försäkringsperiod, antingen samma eller en förändrad premie. Vid annullation så har inte en ny premie erbjudits och därför ses premieförändringen som noll vid annullation. En observation av en försäkring som avslutas genom annullation ger oss information om att den avslutades trots att premieförändringen var noll. I regel får en försäkring inte avslutas under försäkringstiden om inte försäkringsbehovet har upphört eller om så är avtalat, men vissa undantag från detta görs ändå. Vi väljer i detta arbete att inte studera observationer från försäkringar som har annullerats, eftersom vi inte vill råka fånga upp felaktiga effekter från observationer som har avslutats just för att försäkringsbehovet har försvunnit. Förnyelsegraden som därmed modelleras är den andelen av försäkringar som vid försäkringsperiodens slut väljer att förnya sin försäkring, och inte hur många av de som vid ingående försäkringsperiod också finns kvar i nästan försäkringsperiod.

Denna uppdelningen av olika sätt att avsluta sin försäkring på och hur det kan hanteras diskuteras i *Estimating Insurance Attrition Using Survival Analysis* (Fu & Wang, 2015). Fu och Wang uppmärksammar begränsningarna som finns i att låta responsvariabeln anta binära värden (1 för förnyelse, 0 för ickeförnyelse) i logistisk regression som inte låter skilja på om det är en annullation eller ickeförnyelse. En utveckling av detta arbete skulle kunna vara att följa Fu och Wang exempel och istället tillämpa en överlevnadsmodell som använder försäkringens duration i antalet månader som responsvariabel och på så vis kunna se en säsongseffekt av förnyelsegraden och vilka variabler som påverkar när en försäkring kommer att avslutas.

4.5 Obalanserat data

Vårt data är obalanserat när det kommer till andelen observationer av förnyelser och ickeförnyelser. Ett data kallas obalanserat om en klass är överrepresenterad. Vi har sammanlagt 1 247 327 observationer, men endast 24 169 av dessa är ickeförnyelser. Det motsvarar ca 2% av det totala datamaterialet. Eftersom det är en sådan hög andel som väljer att förnya sin försäkring så är det svåra att urskilja

de som inte väljer att förnya. Denna obalans kan leda till att modeller konstrueras som ser ut att ha hög prediktionsförmåga, men det är för att modellerna är bra på att prediktera när förnyelse kommer att ske, inte när ickeförnyelse kommer att ske.

Det finns flera metoder för att hantera obalanserat data, som t.ex. *over sampling* och *under sampling* (ibland även kallad *down sampling*). I over sampling replikerar vi observationer från minoritetsgruppen tills vi har lika många eller önskad andel av minoritets- och majoritetsgruppen. I under sampling behåller vi antalet observationer i minoritetsgruppen och drar istället observationer från majoritetsgruppen tills vi har önskad andel av minoritets- och majoritetsgruppen. Vilken metod som är mest lämplig beror på frågeställning och data (Liu *et al.* , 2006, s. 15-16).

Eftersom vi har en stor mängd observationer så väljer vi att tillämpa under sampling för att på så vis minska tidsåtgången vid modelleringen. En nackdel med denna metod är att information kan förloras i och med reduceringen i antalet observationer och inom olika variabler, men vinsten kan bli en bättre prediktionsförmåga för minoritetsgruppen. Ett problem som kan uppstå i algoritmen Random Forest när obalanserat data används för att konstruera klassifikationsträd är att det i bootstrapsteget av data inte väljs några observationer som är ickeförnyelser, eftersom den andelen är så liten. Algoritmen Random Forest tar även väldigt lång tid på sig vid anpassning på stora data set, medan detta inte är ett lika stort problem i logistisk regression.

I under sampling väljer vi att dra så många observationer från majoritetsgruppen (förnyelse), utan återläggning, så att minoritetsgruppen (ickeförnyelse) kommer utgöra 20% av fullständigt data. Alla observationer som består av ickeförnyelser behålls alltså, men endast en viss del av observationerna som består av förnyelser. Detta för att ickeförnyelserna ska utgöra en större andel av data som modellerna anpassas på. Anledningen till varför vi väljer att ickeförnyelser ska utgöra 20% av data utgår från en analys som presenteras i Avsnitt 6.1.

I Avsnitt 4.3 diskuteras eventuella problem som kan uppstå i en GLM när observationer är beroende, vilket kan antas att observationer av samma försäkringskontrakt men för olika försäkringsperioder är. Denna slumpmässiga reducering av data vad gäller observationer av förnyelser kan hjälpa oss att reducera beroende i data. Men observera att samma försäkringskontrakt trots allt kan förkomma flera gånger i vårt data.

När vi använder oss av under sampling så kommer sannolikheten för förnyelse att

påverkas, eftersom vi manuellt har justerat andelen observerade förnyelser. Skattad sannolikhet för förnyelse kommer alltså att vara lägre i vår modell än i verkligheten. I en logistisk regressionsmodell så är det interceptet som kommer att vara felskattat för att passa vårt fullständiga data. Vi kan därför justera interceptet som skattas när balanserat data används, en justering som kallas *prior correction*.

I prior correction antas att vi vet den sanna andelen av förnyelser i data, τ. I vårt fall så antar vi att vårt fullständiga data utgör sanningen av andelen förnyelser och ickeförnyelser, så vi antar att $\tau = 0.9805$. Andelen av förnyelser i data som används för att skatta modellen är $\bar{y} = 0.80$. Om $\hat{\beta}_0$ är interceptet skattat i modellen så blir då ett intercept som är justerat för att balanserat data används

$$\beta_0 = \hat{\beta}_0 - \log\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right].$$

Observera att om $\tau = \bar{y}$ så är $\beta_0 = \hat{\beta}_0$ (King & Zeng, 2001, s. 144).

I Random Forest kommer sannolikheter för förnyelse att skattas baserat på majoritetsrösten i de olika löven och sedan majoritetsrösten mellan beslutsträden. Samma sak gäller där, att nu baseras majoritetsrösten på mycket färre antal observationer av förnyelse och därför kommer sannolikheterna att ha skattats fel. Vi vill därför rätta till denna felskattning även för Random Forest-modellen.

Låt oss kalla vårt fullständiga observerade data för $(Y, X)_F$ och data som har tagits fram med under sampling för $(Y, X)_U$. För de två dataseten gäller alltså att $(Y, X)_U \subset (Y, X)_F$. Låt oss introducera den binära variabeln s som är 1 om en observation finns i $(Y, X)_U$ och 0 annars. Det antas att utfallet av s inte beror på X, alltså värdet på våra förklarande variabel. Vi väljer slumpmässigt ut vilka av observationerna som representerar en förnyelse som ska ingå i $(Y, X)_U$. Det gör att $P(s|y, x) = P(s|y)$. Vi kan då med Bayes regel skriva

$$P(y = 0|x, s = 1) = \frac{P(s = 1|y = 0)P(y = 0|x)}{P(s = 1|y = 0)P(y = 0|x) + P(s = 1|y = 1)P(y = 1|x)}.$$

Vi vet att $P(s = 1|y = 0) = 1$ eftersom $y = 0$ innebär ickeförnyelse och vi har kvar alla observationer av ickeförnyelse i vårt undersamplade data. Då fås

$$P(y = 0|x, s = 1) = \frac{P(y = 0|x)}{P(y = 0|x) + P(s = 1|y = 1)P(y = 1|x)} = \frac{z}{z + \Omega(1 - z)} = q_U$$

Termen Ω är alltså sannolikheten att en observation av förnyelse finns i undersamplat data. Vi kommer att ta ut 7,95% av förnyelseobservationerna för att ickeförnyelserna ska utgöra 20% av undersamplat data. Då kan ekvationen ovan

skrivas om så att vi får $z = P(y = 0|x)$, som är sannolikheten för ickeförnyelse i det fullständiga data, som en ekvation av q_U, som är sannolikheten för ickeförnyelse i den skattade modellen baserat på under samplat data. Då fås den justerade sannolikheten som

$$z = \frac{\Omega q_U}{\Omega q_U + (1 - q_U)}$$

(Dal Pozzolo *et al.* , 2015).

4.6 Modellering av förnyelsegrad med logistisk regression

4.6.1 Modellen

Vi vill konstruera en logistisk regressionsmodell med två möjliga utfall för responsvariabeln. Responsvariabeln betecknas som Y_i för försök i. Vi låter i modelleringen responsvariabeln anta värden enligt

$$Y_i = \begin{cases} 1 \text{ om förnyelse sker} \\ 0 \text{ om ickeförnyelse sker} \end{cases}$$

och låt utfallet 0 (ickeförnyelse) vara basvärdet i denna modell. Vi har då

$$\eta_i(\beta, \mathbf{x}_i) = \log\left(\frac{P(Y = 1|\mathbf{x}_i)}{P(Y = 0|\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{1,i} + ... + \beta_p x_{p,i} = \mathbf{x}_i^T \boldsymbol{\beta}. \tag{4.2}$$

Om de förklarande variablerna alla hade varit numeriska eller om de hade varit kategoriska med två kategorier så hade p varit antalet förklarande variabler. Nu har vi initialt 5 antal kategoriska förklarande variabler och 11 antal numeriska förklarande variabler. De numeriska förklarande variablerna bidrar alla med en β-parameter var, medan den kategoriska variabel, med index j, med k_j antal kategorier kommer att bidra med $k_j - 1$ antal β-parametrar och den tillhörande variabeln x kommer att vara en dummyvariabel.

Om vi till exempel studerar en variabel vi kallar `geo` som beskriver om en kund bor i en storstad (0), tätort (1) eller landsbygd (2). Det är en kategorisk variabel som har tre kategorier, $k_{geo} = 3$. Denna variabel kommer att bidra med $k_{geo} - 1 = 2$ β-parametrar som vi för tillfället kallar $\beta_{geo,0}$ och $\beta_{geo,1}$. Låt landsbygd vara baskategorin och därför kommer de tillhörande x-värdena att vara

$$\begin{array}{lllll} \text{Om vi studerar en kund i} & \text{storstad} & \rightarrow & x_{geo,0} = 1, \, x_{geo,1} = 0 \\ & \text{tätort} & \rightarrow & x_{geo,0} = 0, \, x_{geo,1} = 1 \\ & \text{landsbygd} & \rightarrow & x_{geo,0} = 0, \, x_{geo,1} = 0. \end{array}$$

På så vis fås tre fall för de tre kategorierna, precis som vi vill, men reducerar antalet parametrar genom att sätta en av kategorierna som baskategori.

Vi har både förklarande variabler som skulle kunna hanteras som kontinuerliga, så som ålder och kundduration, och variabler som naturligt är kategoriska, så som geografiska måttet och bolag. Vi kommer under modelleringens gång att testa att antingen fortsätta att ha vissa variabler kontinuerliga eller dela upp dem i intervall och på så vis istället göra dem kategoriska. Intervallängderna kommer också under modelleringens gång att justeras för att försöka fånga upp effekter som variablerna har på förnyelsegraden och också för att förenkla modellen där fler intervall är överflödiga.

Utifrån modellen fås också de två sannolikheterna

$$\pi_0(\mathbf{x}_i) = P(Y_i = 0|\mathbf{x}_i) = \frac{1}{1 + e^{g(\mathbf{x}_i)}}$$

$$\pi_1(\mathbf{x}_i) = P(Y_i = 1|\mathbf{x}_i) = \frac{e^{g(\mathbf{x}_i)}}{1 + e^{g(\mathbf{x}_i)}}$$

4.6.2 Tolkning av parametrar i logistisk regression

Så vad säger de skattade β-parametrarna oss i en logistisk regressionsmodell? Funktionen i ekvation 4.2 beskriver alltså logoddset för att en förnyelse ska ske och då fås att

$$e^{g(\mathbf{x})} = \frac{P(Y = 1|\mathbf{x})}{P(Y = 0|\mathbf{x})} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} = e^{\mathbf{x}^T\boldsymbol{\beta}}$$

beskriver oddset. Som exempel, säg att x_1 här representerar åldersvariabeln. Vi kan då formulera en oddskvot mellan två olika åldrar, givet alla andra variabler, som

$$\frac{\left(\frac{P(Y=1|\mathbf{x},x_1=21)}{P(Y=0|\mathbf{x},x_1=21)}\right)}{\left(\frac{P(Y=1|\mathbf{x},x_1=20)}{P(Y=0|\mathbf{x},x_1=20)}\right)} = \frac{e^{\beta_0+\beta_1\cdot 21+\ldots+\beta_p x_p}}{e^{\beta_0+\beta_1\cdot 20+\ldots+\beta_p x_p}} = e^{\beta_1}.$$

Vi ser då att e^{β_1} blir oddskvoten för förnyelse vid en åldersökning på 1 år. Parametern β_1 är därmed log oddskvoten för förnyelse vid en åldersökning på 1 år. På samma sätt kan en parameter tolkas som representerar kategoriska variabler, t.ex. kön. Säg att x_2 är 1 om kunden är man och 0 om kunden är kvinna. Oddskvoten för förnyelse mellan man och kvinna blir

$$\frac{\left(\frac{P(Y=1|\mathbf{x},x_2=1)}{P(Y=0|\mathbf{x},x_1=1)}\right)}{\left(\frac{P(Y=1|\mathbf{x},x_2=0)}{P(Y=0|\mathbf{x},x_2=0)}\right)} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2\cdot 1+\ldots+\beta_p x_p}}{e^{\beta_0+\beta_1 x_1+\beta_2\cdot 0+\ldots+\beta_p x_p}} = e^{\beta_2}.$$

4.6.3 Modellkonstruktion

För att anpassa en modell till vårt data så väljer vi att använda allt data fram till 2015. På så vis kan sedan modellens prediktionsförmåga testas på observationer från 2016, så kallat *out of sample*-test. Det är alltså på data fram till och med 2015 som under sampling tillämpas (se Avsnitt 4.5) och som blir vårt *träningsdata*, medan data från 2016 blir vårt *testdata*. Vårt träningsdata består av 118 575 balanserade observationer och testdata av 31 042 antal observationer.

Istället för att börja med att inkludera alla förklarande variabler samtidigt så byggs en univariat logistisk regressionsmodell med respektive förklarande variabel som enda förklarande variabel och med förnyelse som responsvariabel. För varje sådan modell beräknas en likelihoodkvotstatistika som testar hypotesen om den förklarande variabeln bör inkluderas i modellen eller inte. Vi kommer alltså att basera testet på loglikelihoodvärdet av en modell med endast en interceptparameter, modellen L_0, mot en där vi också har inkluderat parametrar för olika värden på inkluderad förklarande variabel, modellen L_1. Likelihoodkvotstatistikan blir då

$$G = -2\dot{(}L_0 - L_1).$$

Statistikan är χ^2-fördelade med antal frihetsgrader lika med skillnaden i antalet skattade parametrar i de två modellerna (Agresti, 2012, s. 12). För att avgöra om vi ska fortsätta att använda kontinuerliga variabler som just kontinuerliga så görs även en grafisk analys av dem och testar att inkludera dem som kontinuerliga eller kategoriska i sina univariata modeller.

I Tabell 2 kan likelihoodkvotstatistikan ses och tillhörande p-värde för varje förklarande variabel. Som vi nämnde tidigare så pseudonymiseras variablerna på grund av sekretesskäl. Det vanligaste är att man använder en signifikansnivå på 10% eller 5% när man väljer förklarande variabler i en modell, men här väljs ett något högre kriterium på 20% initialt i denna univariata analys för att inte råka exkludera en variabel som skulle ha en inverkan i en större modell. Vi ser där att alla förklarande variabler blev signifikanta i ett LRT-test utom `variabel8` som representerar systemen som observationerna har hanterats i. Att inte `variabel8` är signifikant är bra, då det indikerar att försäkringarna har hanterats likvärdigt för våra syften i det här arbetet. Alla förklarande variabler inkluderas alltså i en större logistisk regressionsmodell bortsett från `variabel8`.

Variabel	**LRT**	*df*	**p-värde**
variabel1	14.4	1	< 0.001 *
variabel2	178.95	1	< 0.001 *
variabel3	974.9	1	< 0.001 *
variabel4	2.7	1	0.1030 *
variabel5	229.9	1	< 0.001 *
variabel6	16.5	6	0.0110 *
variabel7	270.9	6	< 0.001 *
variabel8	0.3	1	0.6058
variabel9	2582.8	11	< 0.001 *
variabel10	3500.6	1	< 0.001 *
variabel11	6150.4	1	< 0.001 *
variabel12	4.0	1	0.0466 *
variabel13	968.0	1	< 0.001 *
variabel14	146.56	3	< 0.001 *
variabel15	5484.3	1	< 0.001 *
variabel16	127.0	1	< 0.001 *
variabel17	1477.9	5	< 0.001 *
variabel18	830.9	2	< 0.001 *

Tabell 2: *Likelihoodkvottest av varje förklarande variabel i en univariat logistisk regressionsmodell. Signifikanta variabler på 20%-nivån är markerade med *.*

Vi upptäcker en viss problematik med att inkludera variablerna gällande antal skador och medelskadekostnaden tillsammans i en större modell. Det är förståeligt att det uppstår viss problematik, eftersom de är beroende i viss mening. Om kunden aldrig har haft några skador så kommer antalet skador också alltid att vara noll och även medelskadekostnaden. Det motsatta gäller dock inte, eftersom skador kan ha rapporterats där det inte har betalats ut någon ersättning.

Efter några försök med kombinationer av dessa variabler och även kombinationer där vi låter variablerna vara kontinuerliga eller anta kategoriska värden baserat på intervall istället så landar vi i en modell där antal skador har inkluderats som kontinuerlig och medelskadekostnad under försäkringsperioden med endast två kategorier, bestående av "medelskada $= 0$" och "medelskada $\neq 0$". I Figur 4 kan förnyelsegraden ses för aggregerat data för några intervall av medelskador och vi tycker oss kunna se en svag ökning för ackumulerad medelskada under hela försäkringstiden (röda trianglar), men eventuellt en mer slumpmässighet för medelskada under gällande försäkringsperiod (svarta punkter). Eftersom en sådan

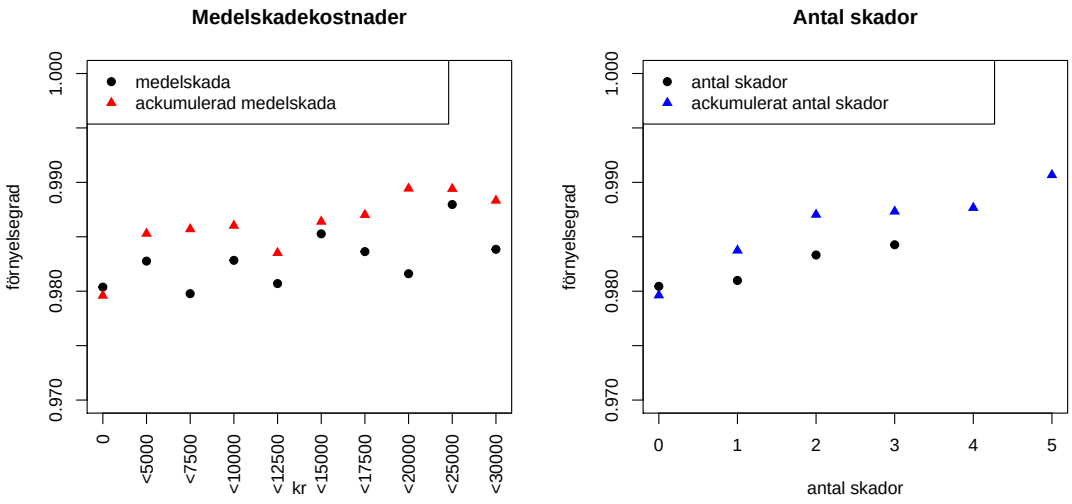
uppdelningen på bara två kategorier av medelskadekostnaden i modellen endast säger om det har skett en skada där ersättning har utbetalats eller inte, så väljer vi att definiera en ny kategorisk förklarande variabel som kombinerar medelskadekostnad och antalet skador. Vi kallar den för tillfället `skada` och definierar dess kategorier som

$$\texttt{skada} = \begin{cases} 0 \text{ om skada inte har inträffat,} \\ 1 \text{ om skada någonsin har inträffat och ingen ersättning har betalats,} \\ 2 \text{ om skada någonsin har inträffat och ersättning har betalats.} \end{cases}$$

Denna variabel inkluderas istället för antal skador och medelskadekostnad i vår stora logistiska regressionsmodell. Vi provar även att utöka variabeln `skador` så att någon kategori representerar en ökning av skador som har fått ersättning, eftersom ackumulerat antal skador trots allt verkade ha en effekt i modellen. I Figur 4 ses en ökning av förnyelsegraden i data som är aggregerat för olika antal skador under försäkringstiden (blå trianglar). Vi kan ju även se något av en ökning i förnyelsegraden för antal skador som har skett under gällande försäkringsperiod (svarta punkter). Men en sådan utökning av kategorierna i `skada` blir inte signifikant och inte heller kategorin 2 i variabeln. Kategorierna 0 och 2 slås därför ihop och får nedan indelning av variabeln i vår stora logistiska regressionsmodell.

$$\texttt{skada} = \begin{cases} 0 & \text{om skada inte har inträffat eller} \\ & \text{skada har inträffat och ersättning har betalats,} \\ 1 & \text{om skada någonsin har inträffat och ingen ersättning har betalats.} \end{cases}$$

Vi anonymiserar även denna variabel och kallar den därför inte längre `skada`. Den blir i en univariat logistisk regressionsmodell signifikant och finns redan representerad i Tabell 2. Utöver detta så exkluderas ingen annan variabel, men vi justerar intervallgränserna på vissa variabler och slår även ihop några intervall helt.



Figur 4: *Förnyelsegrad baserad på aggregerat data för antal skador och medelskadekostnader. Datat har inte justerats för obalans i förnyelse och ickeförnyelse.*

I Tabell 3 kan vi se beräknade GVIF och justerat GVIF för att kunna jämföra måtten för variabler med olika dimensioner. Ingen av de förklarande variablerna ger upphov till justerade GVIF-värden som når upp till 5, vilket indikerar att det inte finns upphov till multikolinjäritet i vår modell.

Variabel	$GVIF$	df	$GVIF^{1/2df}$
`variabel1`	1.224	1	1.106
`variabel2`	1.085	1	1.042
`variabel3`	1.104	1	1.051
`variabel9`	1.382	11	1.015
`variabel10`	1.737	2	1.148
`variabel11`	1.577	1	1.256
`variabel12`	1.007	1	1.003
`variabel13`	1.163	1	1.078
`variabel14`	1.020	2	1.005
`variabel15`	1.167	1	1.080
`variabel16`	1.061	1	1.030
`variabel17`	1.036	3	1.030
`variabel18`	1.207	1	1.099

Tabell 3: *GVIF för förklarande variabler i logistisk regressionsmodell.*

4.7 Modellering av förnyelsegrad med Random Forest

Precis som vid modelleringen i en logistisk regressionsmodell så anpassas vårt klassifikationsträd på allt data från 2004-2015 som vi har undersamplat, vårt *träningsdata*, som består av 118 575 antal observationer. Modellen testas sedan på data från 2016, vårt *testdata.* Observera att det finns en viss slumpmässighet i konstruktionen av ett beslutsträd med algoritmen Random Forest då vi för varje beslutsträd som konstrueras slumpar fram ett dataset. I varje split slumpas också fram vilka förklarande variabler vi kan välja bland att göra en uppdelning i, alltså var det ska placeras ut en gren. Som det har rekommenderats så väljs $\sqrt{p} = m$ antal variabler i varje steg som en uppdelningen kan ske vid där p är antalet förklarande variabler.

Vi inkludera förklarande variabler som både är kategoriska och numeriska och vi kan även här välja om vi istället vill att numeriska variabler ska delas in i intervall och representeras som kategoriska. Alla förklarande variabler inkluderas och vi låter de som naturligt kan ses som numeriska vara just numeriska och annars antar de kategoriska värden.

Vi har ett stort dataset att jobba med och väljer därför att begränsa algoritmen genom att sätta lövstorleken till 100 och stoppar då alltså algoritmen när löven innehåller 100 observationer. Storleken på datasetet som slumpas fram sätts till 2/3 av vårt data. Vi börjar med att konstruera 1000 träd och ser att OOB-felskattningen ser ut att konvergera runt ca 300 träd, så för att minska tidsåtgången vid modelleringen så väljs därför att göra 300 träd. De kommer alla att ha 1186 löv.

I konstruktionen av klassifikationsträd lägger man inte lika stor vikt vid att analysera vilka variabler som ska inkluderas eller exkluderas ur modellen, som det ofta gör vid generaliserade linjära modeller. Om vi ser att vissa variabler inte verkar ha så stor inverkan på modellens Gini-index eller exakthet i klassifikation, så skulle de eventuellt kunna exkluderas ur modellen. Men om det inte heller tros att de leder till ökad felklassificering så skadar det inte att låta dem vara kvar.

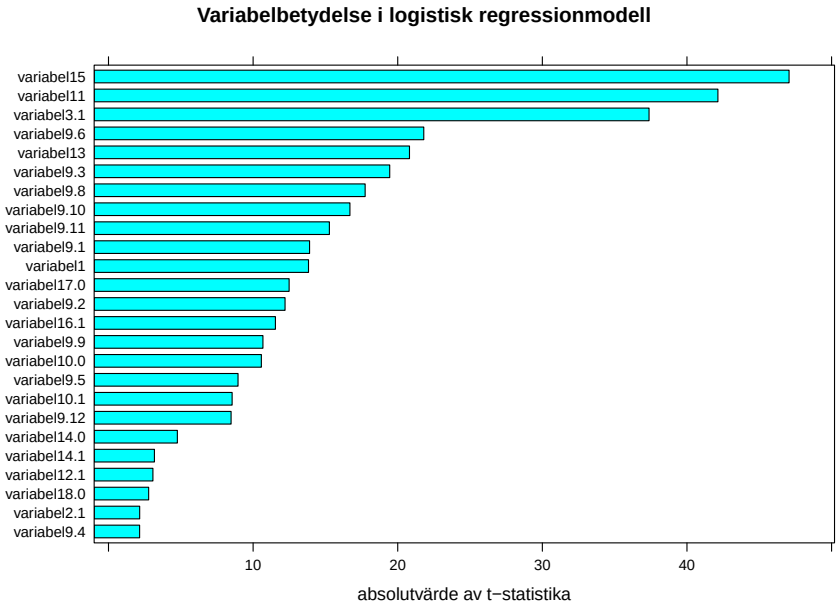
5 Resultat

I detta avsnitt presenteras resultat baserat på modellkonstruktionerna av vår logistiska regressionsmodell och Random Forest-modell.

5.1 Variabler i logistisk regression

I den slutgiltiga logistiska regressionsmodellen inkluderades 13 förklarande variabler varav 4 var kontinuerliga medan 9 var kategoriska, vilket gav upphov till 26 skattade parametrar. De presenteras alla i Tabell 8 i Appendix B, avrundat till 2 decimaler.

I Figur 5 kan vi se absolutvärdet av t-statistikan för alla parametrar i vår slutgiltiga modell, vilket kan ge en indikation på vilken betydelse de olika förklarande variablerna har haft för modellen. Att det står t.ex. `variabel3.1` betyder alltså `variabel3` med kategori 1 och `variabel9.6` är då `variabel9` med kategori 6. Vi ser där att `variabel15`, `variabel11` och `variabel3` har störst t-statistika.

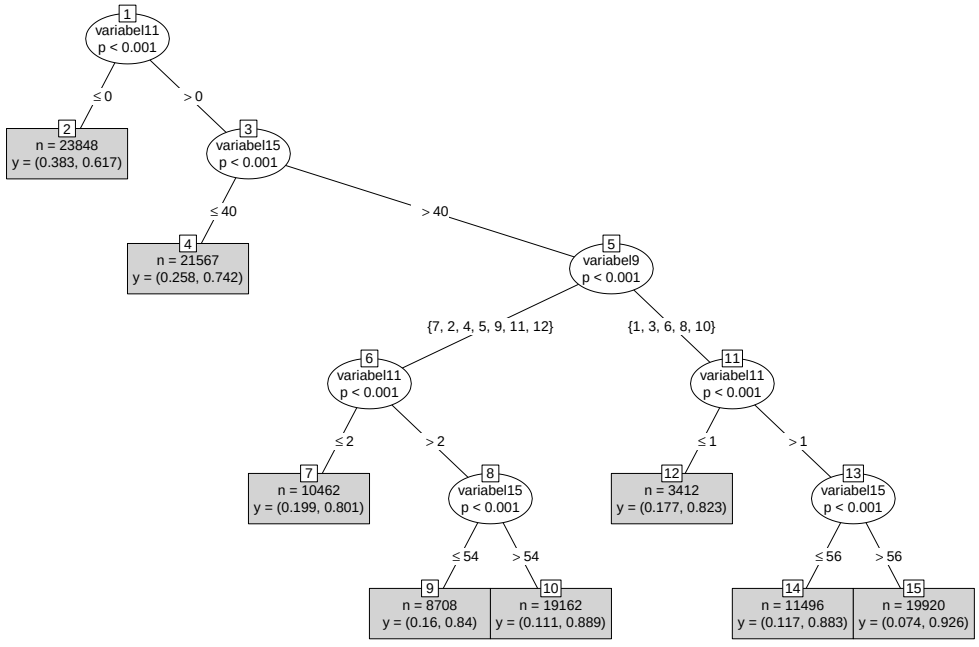


Figur 5: *Variabelbetydelse i logistisk regressionsmodell baserat på absolutvärdet av t-statistikan.*

5.2 Variabler i Random Forest

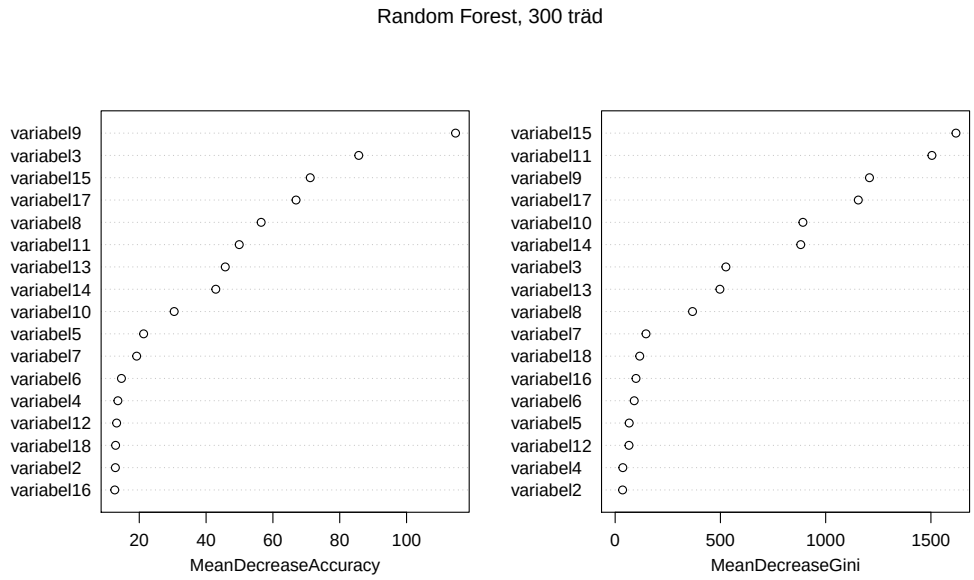
Eftersom beslutsträd är en icke-parametrisk metod för att prediktera utfallet på en variabel kommer det inte att kunna göras samma representation av de förklarande variablerna som vi tar med i modellen, alltså i termer av odds eller skattade parametrar, som vi kan göra i logistisk regression. Särskilt inte när vi använder oss av Random Forest där vi bygger 300 beslutsträd med vardera 1186 löv för att reducera variansen i prediktionerna. En sådan analys kan lättast göras för mindre träd där vi inte tillämpar bagging som vi gör för Random Forest.

För att få en illustration av detta så skapas ett enkelt beslutsträd där vi säger att algoritmen ska stanna när löven vardera innehåller mindre än 27 500 observationer som kan ses i Figur 6. Detta träd har alltså 8 löv, vilket då kan jämföras med ett av våra träd som har 1186 löv. I ett sådant här mindre träd kan sannolikheterna studeras närmare för att se om besluten i trädet leder till högre eller lägre sannolikhet för förnyelse. Dessa skattade sannolikheter ses i varje löv som $y =$ (sannolikheten för ickeförnyelse, sannolikheten för förnyelse). I denna illustration är dessa sannolikheter inte justerade för att passa ett obalanserat data utan de baseras på det balanserade data som använts för att anpassa trädet.



Figur 6: *Beslutsträd för illustration.*

Ett resultat från Random Forest-modellen som byggs med 300 träd är hur viktiga variablerna är för modellen. I Figur 7 kan vi se hur de olika förklarande variablerna har minskat *Accuracy* och *Gini-index* i konstruktionen (se Avsnitt 2.2.2 och 2.2.4). MeanDecreaseAccuracy visar hur mycket exaktheten vid klassifikation minskar vid exkluderingen av en variabel. Vi ser att exaktheten minskar som mest vid exkludering av `variabel9`. MeanDecreaseGini visar hur mycket Gini-index har minskat vid grenar placerade i de olika förklarande variablerna. Gini-index har minskat mest när grenar har placerats i variabeln `variabel15`. Variabeln `variabel15` visar sig även vara viktig i MeanDecreaseAccuracy och vice versa för `variabel9`.



Figur 7: *Medelminskning i exakthet (accuracy) samt medelminskning i Gini-index i Random Forest algoritm med 300 träd.*

5.3 Prediktionsförmåga

Träningsdata

När modellernas prediktionsförmåga testas börjar vi med att titta på hur modellerna predikterar utfallen för träningsdata, alltså observationerna vi har använt för att anpassa modellerna. För denna prediktion justeras inte sannolikheterna som modellen ger oss, eftersom träningsdata är balanserat och modellerna är anpassade för just denna andel av förnyelser och ickeförnyelser.

För att kunna göra en prediktion så måste ett tröskelvärde, c, bestämmas som gör så att vi predikterar $Y = 1$ när $P(Y = 1) \geq c$. Vad vi då gör är att ett predik-

tionsmått används som kan definieras av termerna i klassifikationstabellen i Avsnitt 2.3. Vi är främst intresserade av prediktionsförmågan gällande ickeförnyelse, $Y = 0$, och vill därför att måttet $TNR = TN/N$ (*True Negative Rate*) ska vara så högt som möjligt samtidigt som $FNR = FN/P$ (*False Negative Rate*) ska vara så lågt som möjligt. Vi kan beräkna dessa mått för olika trösklar och sedan välja den tröskeln som maximerar summan $TNR + (1 - FNR)$. I den logistiska regressionsmodellen så beräknas detta maximum för tröskelvärdet $c = 0.8081$ och för Random Forest till 0.9833 för träningsdata.

I Tabell 4 kan vi se att av de som faktiskt har utfallet förnyelse ($Y = 1$) i den logistiska regressionsmodellen så predikterades 66% av dessa som förnyelser, samt så predikterades 68% av ickeförnyelserna rätt. Det ger en total prediktionsgrad på 66% ($(TP+TN)/N_p$). Motsvarande andelar för Random Forest-modellen är 70%, 80% och totalt 72%. Så Random Forest-modellen predikterar något bättre för data som modellen är anpassad på än vad den logistiska regressionsmodellen gör.

Logistisk regression, träningsdata

		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.66 (62 168)	0.32 (7 622)
	$Y = 0$	0.34 (32 692)	0.68 (16 093)
	Totalt antal	94 860	23 715

Random Forest, träningsdata

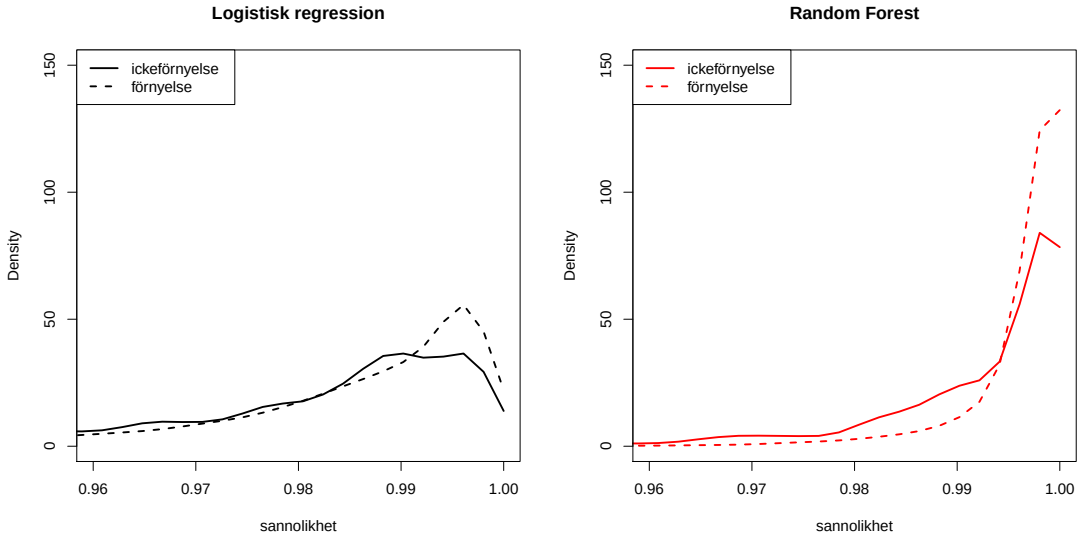
		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.70 (66 528)	0.20 (4 643)
	$Y = 0$	0.30 (28 332)	0.80 (19 072)
	Totalt antal	94 860	23 715

Tabell 4: *Klassifikationstabeller över andelen prediktioner i logistisk regressionsmodell och Random Forest för träningsdata. Prediktionen $Y = 1$ görs när $P(Y = 1) \geq 0.8081$ för logistisk regression och när $P(Y = 1) \geq 0.9833$ för Random Forest.*

Testdata

Vi vänder oss nu istället till testdata från 2016 med 31 042 observationer. Nu justeras sannolikheterna för förnyelse i enlighet med Avsnitt 4.5 och vi kan i Figur 8 se fördelningen över de skattade sannolikheterna för förnyelse. Denna figur ska illustrera om det kan ses någon skillnad på skattade sannolikheter för observationer i testdata som vi vet har utfallet förnyelse (streckade linjer) samt ickeförnyelse (heldragna linjer). Som vi ser så är skillnad på de skattade sannolikheterna för observerat data med förnyelse och ickeförnyelse inte jättestor. Det kan ses något av en förskjutning till vänster av sannolikheterna för ickeförnyelse och den förskjutningen är lite större i Random Forest-modellen som illustreras till höger. Vi kan också se att för förnyelserna är en större andel av sannolikheterna väldigt höga. Här kan vi

då också se problemet med att använda ett gemensamt tröskelvärde eller en låg tröskel på t.ex. $c = 0.5$ för att göra en prediktion av förnyelse eller ickeförnyelse. Eftersom ingen sannolikhet är lägre än 0.5 så kommer då alla observationer predikteras som förnyelse. Det kan därför vara bättre att använda en annan tröskel än 0.5.



Figur 8: *Fördelningen av skattade sannolikheter för förnyelse baserat på testdata från 2016 i logistisk regressionsmodell och Random Forest. De heldragna linjerna är skattade sannolikheter för förnyelse för observationer som har utfallet ickeförnyelse. De streckade linjerna är skattade sannolikheter för förnyelse för observationer med utfallet förnyelse.*

I Tabell 5 kan vi se klassifikationstabeller för prediktioner gjorda med våra modeller på testdata från 2016. Tröskelvärdena vi nu använder är $c = 0.9910$ för logistisk regression och $c = 0.9967$ för Random Forest. Den totala prediktionsgraden är 44% för den logistiska regressionsmodellen och 77% för Random Forest.

Logistisk regression, testdata

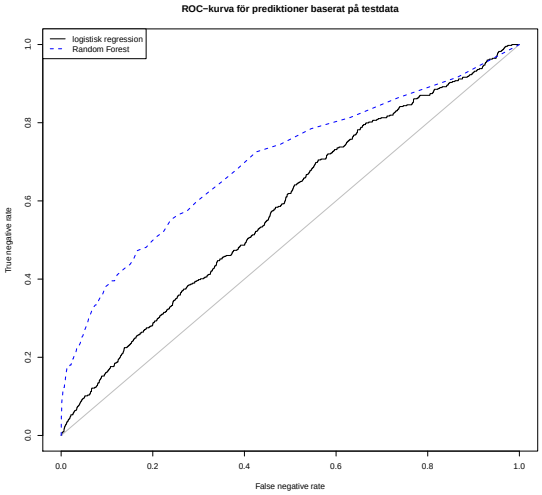
		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.44 (13 433)	0.30 (134)
	$Y = 0$	0.56 (17 155)	0.70 (320)
Totalt antal		30 588	454

Random Forest, testdata

		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.76 (23 270)	0.45 (204)
	$Y = 0$	0.24 (7 318)	0.55 (250)
Totalt antal		30 588	454

Tabell 5: *Klassifikationstabeller över andelen prediktioner i logistisk regressionsmodell och Random Forest för testdata. Prediktionen* $Y = 1$ *görs när* $P(Y = 1) \geq 0.9910$ *för logistisk regression och när* $P(Y = 1) \geq 0.9967$ *för Random Forest.*

I Figur 9 kan vi se ROC-kurvan för prediktionsmåtten True Negative Rate och False Negative Rate för alla tröskelvärden mellan 0 och 1 för prediktion. En större yta under kurvan tyder på att modellen är bättre på att skilja ett negativt utfall från ett positivt. Vi ser där att Random Forest har betydligt större yta under kurvan, som uppgår till 0.70. Ytan under ROC-kurvan tillhörande logistisk regression är 0.58.



Figur 9: *ROC-kurva för True Negative Rate och False Negative Rate i logistisk regressionsmodell och Random Forest för testdata.*

6 Analys

I detta avsnitt analyseras den logistiska regressionsmodellen och Random Forest-modellen som konstrueras i Avsnitt 4. Vidare presenteras analysen bakom valet av andelen som ickeförnyelser ska utgöra i det data som används för modelleringen.

6.1 Andelen i balanserat data

För att ta reda på hur vårt data ska balanseras, alltså vilka andelar vi vill att förnyelseobservationer och ickeförnyelseobservationer ska ha, så studeras ROC-kurvor och hur stor ytan under dem är, för att avgöra vilken andel som är lämplig att använda. När antalet observationer reduceras i data riskerar vi att förlora värdefull information och vi vill studera om det händer när vi balanserar vårt data och anpassar modellerna på det. Förhoppningen är att modeller anpassade på balanserat data ska ge oss bättre resultat än för modeller anpassade på obalanserat data, eftersom vi låter ickeförnyelseobservationerna få en större vikt i data.

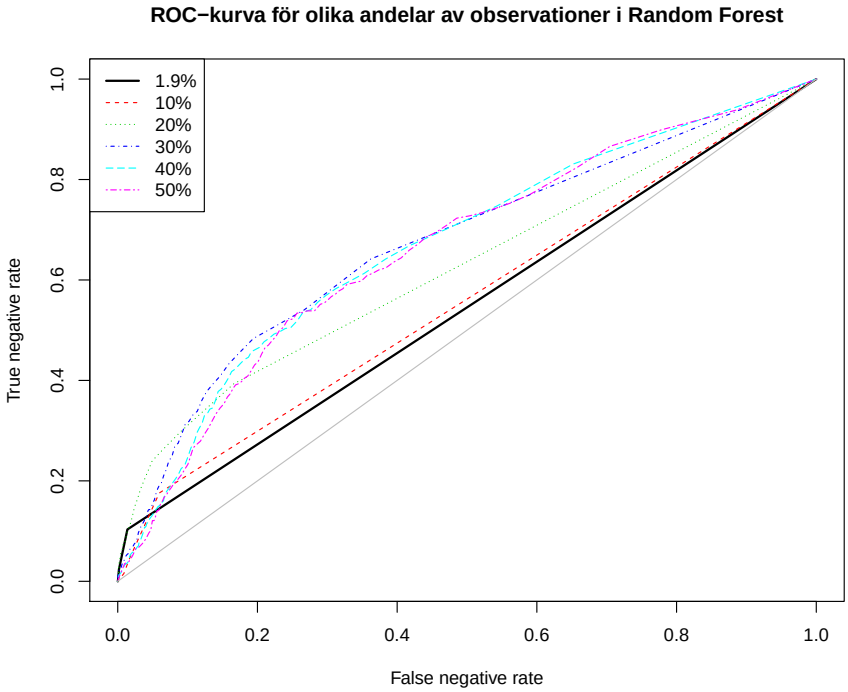
Logistisk regression

Vi använder alltså den logistiska regressionsmodellen vi kommer fram till i Avsnitt 4.6.3 och anpassar den på data med olika andelar av förnyelseobservationer, men behåller alltid alla ickeförnyelseobservationer. Getta görs för originaldata så att ickeförnyelseobservationerna utgör 1.9%, och sedan så att ickeförnyelseobservationerna utgör 10%, 20%, 30%, 40% och 50%. För varje dataset så ritas en ROC-kurva och vi beräknar tillhörande yta under kurvan, som vi vill ska vara så stor som möjligt. ROC-kurvorna för de olika modellerna blir så snarlika att ytan under dem endast skiljer sig med tusendelar. Det indikerar alltså att balanseringen av data i vår logistiska regressionsmodell inte förbättrar prediktionen för ickeförnyelser, men inte heller försämras. Vinsten är då tidsåtgången vid anpassningen av modellen, men någon annan effekt av balanserat data uppstår inte. Även om skillnaderna är små så är den andelen som ger upphov till störst yta under ROC-kurvan den där ickeförnyelserna utgör 20% av data och därför använder vi den andelen i vårt balanserade data.

Random Forest

Samma analys kommer inte att presentera för vår Random Forest-modell, då anpassningen av obalanserat data till modellen från Avsnitt 4.7 är för omfattande. Vi presenterar därför en analys på en mindre algoritm istället, för att åtminstone få en uppfattning om hur balanserat data påverkar en Random Forest-modell. Vi begränsar oss till 100 träd och 1 000 observationer per löv och i Figur 10 kan

ROC-kurvor ses för respektive modell. Andelarna som anges är hur mycket ickeförnyelserna utgör av data som använts för att anpassa modellerna. Ytan under ROC-kurvorna ökar något med andelen ickeförnyelser och antar värdena 0.545, 0.558, 0.625, 0.672, 0.668 och 0.662 för respektive andel i stigande ordning. Även om det är andelen 30% som ger upphov till den största ytan under ROC-kurvan så används data som balanseras så att ickeförnyelserna utgör 20% av data, så att en jämförelse med den logistiska regressionsmodellen lättare kan göras.



Figur 10: *ROC-kruva för True Negative Rate och False Negative Rate i Random Forest-modeller anpassade på olika andelar av ickeförnyelser i data.*

En liknande analys av andelen som ska användas i ett balanserat data vid modellering med logistisk regression och Random Forest har gjorts i artikeln *Storm Prediction: Logistic Regression vs. Random Forest for Unbalanaced Data* (Ruiz-Gazen & Villa, 2007) där mått kallade *False Alarm rate* och *Threat Score* har använts. De måtten är också kvoter baserade på utfallen i en sådan klassifikationstabell som vi har använt oss av i detta arbete. Om vi genomför samma analys med dessa mått istället så får vi liknande resultat vad gäller valet av andelen i balanserat data som vi nyss presenterade för både logistisk regression och Random Forest. De ser även, precis som vi, att prediktionsförmågan varken försämras eller förbättras för logistisk regression när vi balanserar data som modellen anpassas på. De ser

även det för Random Forest-modellen, medan vi ser en antydan till förbättring i prediktionsförmågan, särskilt vad gäller ickeförnyelser, vilket diskuteras i nästa avsnitt.

6.2 Effekten av balanserat data

Som vi kunde se i föregående avsnitt vid valet av andelen i balanserat data så varken förlorar vi någon information eller vinner någon prediktionsförmåga i den logistiska regressionsmodellen. Vad gäller Random Forest ser det dock ut som att prediktionsförmågan har ökat tack vare anpassningen till balanserat data. I Tabell 6 kan vi se effekten som balanserat data har haft på en Random Forest-modell med 100 träd och 1 000 observationer i varje löv. Prediktionsförmågan gällande ickeförnyelser ($Y = 0$) har ökat, men det totala prediktionsmåttet är högre för obalanserat data, nämligen 97%. För balanserat data är det totala prediktionsmåttet 83%.

Anledningen till att den högra tabellen här inte har samma värden som den högra tabellen i Tabell 5 är just för att vi har olika antal träd och observationer i löven för de olika modellerna. En jämförelse mellan dessa kan göras för att se en viss effekt som ökningen av träd och minskningen av observationer i löven gör på prediktionerna.

Random Forest, andel ickeförnyelser = 0.019 (obalanserat)

		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.99 (30 165)	0.90 (407)
	$Y = 0$	0.01 (423)	0.10 (47)
	Totalt antal	30 588	454

Random Forest, andel ickeförnyelser = 0.2 (balanserat)

		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.83 (25 504)	0.61 (276)
	$Y = 0$	0.17 (5 084)	0.39 (178)
	Totalt antal	30 588	454

Tabell 6: *Klassifikationstabeller över andelen prediktioner i Random Forest för testdata, modeller anpassade på obalanserat och balanserat data. Prediktionen $Y = 1$ görs när $P(Y = 1) \geq 1$ för båda modellerna.*

6.3 Prediktion år för år

För att testa ytterligare hur bra modellerna är på att prediktera förnyelser och ickeförnyelser så testar vi hur modellerna predikterar ett år framåt när vi endast har data ett år tillbaka. Vi vet att klassifikationsträd kan vara väldigt känsliga

och förändras mycket även om två dataset är väldigt lika varandra. Även slumpmässigheten i val av dataset och förklarande variabler vid konstruktionen av Random Forest kan bidra till instabilitet i metoden. Steget av bagging i algoritmen hjälper oss med den instabiliteten, men när vi här också utför prediktionerna över flera år får vi också en bild som eventuellt är mer stabil av prediktionsförmågan. Vi kommer därför att konstruera en klassifikationstabell där elementen utgör ett medelvärde av flera prediktioner. Detta görs för data från 2007 till 2016, så alltså 10 sådana här prediktioner. Vi använder data från 2007 för att anpassa vår modell, för att sedan prediktera år 2008, och så fortsätter vi så för alla år och får en klassifikationstabell per år. Sedan tas alltså medelvärdet av elementen i alla dessa tabeller. Alla modeller anpassas på balanserat data där ickeförnyelserna utgör 20% och vi justerar sedan utfallet från modellerna så att vi kan prediktera data som är obalanserat.

I Tabell 7 kan vi se medelvärdet av prediktionerna som görs för 10 år. Random Forest modellerna ser ut att prediktera utfallet förnyelser ($Y = 1$) något bättre än vad de logistiska regressionsmodellerna gör och tvärtom vad gäller att prediktera ickeförnyelserna. De totala prediktionsmåtten är 65% för logistisk regression och 72% för Random Forest.

Logistisk regression, testdata
medelvärdet av prediktioner under 10 år

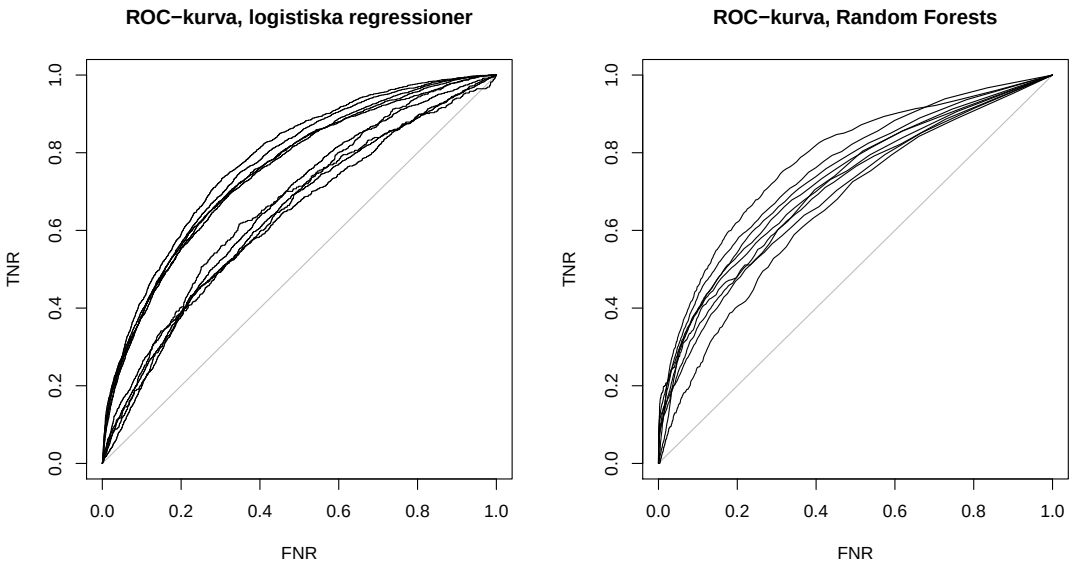
		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.65 (57 250)	0.34 (583)
	$Y = 0$	0.35 (30 648)	0.66 (1 203)
Medelvärde, totalt antal		87 898	1 786

Random Forest, testdata
medelvärdet av prediktioner under 10 år

		Observerat	
		$Y = 1$	$Y = 0$
Predikterat	$Y = 1$	0.72 (63 039)	0.39 (692)
	$Y = 0$	0.28 (24 859)	0.61 (1 094)
Medelvärde, totalt antal		87 898	1 786

Tabell 7: *Medelvärdet av klassifikationstabeller över andelen prediktioner i logistisk regressionsmodell och Random Forest för testdata. Prediktionerna är gjorda ett år framåt och modellerna är anpassade på data från ett år. Detta har gjort för åren 2007 till 2016 (10 prediktioner).*

I Figur 11 kan vi se ROC-kurvor för alla prediktioner för de tio åren. I den vänstra bilden över ROC-kurvorna tillhörande logistiska regressionsmodeller så ser vi ett glapp mellan två grupper av kurvor. De högre kurvorna tillhör tidigare år, medan kurvorna närmast den diagonala linjen är för år närmare 2016. Det samma gäller inte för Random Forest-modellerna där ROC-kurvorna är helt blandade och vi inte kan se något mönster med åren. Den genomsnittliga ytan under kurvorna är 0.70 för logistisk regression och 0.72 för Random Forest.



Figur 11: *ROC-kurva för True Negative Rate och False Negative Rate i logistiska regressionsmodeller och Random Forest-modeller anpassade på data från olika år och prediktioner gjorda för efterföljande år.*

7 Diskussion

I detta avsnitt kommer vi att diskutera de resultat som presenterats i Avsnitt 5 och som vi kommer fram till i analysen av modellerna i Avsnitt 6.

7.1 Modellerna

Vi har i detta arbete anpassat en logistisk regressionsmodell och en Random Forest-modell på observationer över boendeförsäkringar som varit gällande någon gång från 2004 till 2016 hos något bolag inom Dina Försäkringar-federationen. Vi önskar att modellerna ska kunna säga oss någonting om kund- och försäkringsspecifika egenskaper har en inverkan på om en kund väljer att förnya sin försäkring eller inte vid försäkringsperioden slut och i sådana fall vilka egenskaper som har en inverkan. Sannolikheten att en kund väljer att förnya sin försäkring modelleras, som också är en skattning på förnyelsegraden för en sådan kund. De två modellerna som har tillämpats är olika i sin utformning, både vad gäller matematisk anpassning till data, antaganden om samband mellan responsvariabel och förklarande variabler och begränsningar i när de kan användas.

I logistisk regression antar vi att det finns ett linjärt samband mellan logoddset för förnyelse och förklarande variabler. Det antas att responsvariabeln följer en fördelning inom exponentialfamiljen, i vårt fall binomialfördelningen, och vi skattar parametrarna med maximumlikelihoodmetoden. I Random Forest antas inte något form på sambandet över huvud taget, vi antar ingen fördelning för responsvariabeln och använder oss av minimering av ett Gini-index för att skapa homogena löv i modellen. I logistiska regressionsmodellen har vi valt att endast inkludera de förklarande variablerna i linjära termer för de kontinuerliga variablerna och inte som t.ex. kubiska eller kvadratiska polynom. Några försök att inkludera kvadratiska termer gjordes om en kontinuerlig variabel inte blev signifikant, men gav aldrig ett tillfredsställande resultat, så den variabeln fick då fortsätta ha en linjär term eller definieras som kategorisk istället. Vi valde även att lägga vikten vid analys av linjära termer för att kunna göra en lätt tolkning av förändringen i odds mellan olika värden eller kategorier på de förklarande variablerna. Så skillnader i val av förklarande variabler och prediktionsförmåga kan ha uppstått på grund av att modellerna har olika begränsningar. Medan vi endast fångar upp de linjära sambanden i logistisk regression så begränsar vi oss inte alls i Random Forest-modellen, som eventuellt fångar icke-linjära samband som beskriver förnyelsegraden bättre.

Frågan är om de jämförelser vi sedan gör av modellerna blir rättvis när vi begränsar sambanden i logistisk regression, men inte i Random Forest. Hade den

logistisk regressionsmodellen kunnat prediktera ännu bättre om vi gjort en grundligare analys av eventuella kvadratiska och kubiska termer i modellen? Denna analys skulle vara tidsomfattande, särskilt för den mängden variabler vi testar att inkludera i modellen. Detta leder oss till en fördel hos en Random Forest-modell: att tiden det tar för att anpassa en modell är betydligt kortare än för en logistisk regressionsmodell för detta data. Vi ombeds inte heller ta hänsyn till beroende i data, beroende mellan förklarande variabler eller fördelning för responsvariabeln, vilket också minskar tidsåtgången vid modellering med Random Forest.

I Tabell 5 kan vi se resultatet av prediktion gjord för testdata från 2016, för modeller som har anpassats på data från 2004-2015. Modellernas totala prediktionsgrad är 44% för den logistiska regressionsmodellen och 77% för Random Forest-modellen. Baserat på dessa siffror skulle vi kunna säga att Random Forest-modellen predikterar bäst, men det är något missvisande då andelen förnyelser utgör en så stor del av data. Det är ickeförnyelserna som är svårast att fånga upp vid prediktionen och den logistiska regressionsmodellen lyckas prediktera 70% av dessa rätt, medan Random Forest-modellen endast predikterar ca hälften av dessa rätt.

Vi vänder oss sedan till Tabell 7 där vi använder modellerna vi kommer fram till i Avsnitt 4.6.3 och 4.7 för att se hur väl de predikterar förnyelse och ickeförnyelse ett år framåt när vi endast har använt data från ett år tillbaka för att anpassa modellen. Eftersom detta görs för 10 år så minskas variationen i resultatet när vi tar medelvärdet av kvoterna i klassifikationstabellerna. Modellerna visar sig här vara nästan likvärdiga i sin prediktionsförmåga, även om Random Forest-modellen har något högre total prediktionsgrad. Prediktionsförmågan här för ickeförnyelserna är ungefär den samma för logistisk regressions som när anpassningen gjordes på ett större data, medan prediktionsförmågan har ökat för Random Forest-modellen när det kommer till ickeförnyelserna. Detta talar alltså för att vi inte behöver använda allt data vi har för att prediktera ett år framåt, och kan få liknande eller bättre resultat ändå.

7.2 Modellantagande om oberoende observationer

Två antaganden som måste vara uppfyllda för att resultatet och modelleringen av en GLM ska vara tillfredsställande samt genomförbart är att observationerna som används ska vara oberoende och även att de förklarande variablerna ska vara oberoende. Vad gäller oberoende observationer så valde vi att bortse från den eventuella försäkringsspecifika effekt som kan finnas i försäkringarna och leda till en ökad eller minskad sannolikhet för förnyelse som beror på försäkringstagarna i sig

och inte de förklarande variabler som vi använder oss av. Men, vi genomför även ett test av modellerna där data som modellerna anpassas på inte kan ha denna effekt och eventuella beroende, nämligen analysen över tio år där vi predikterar ett år framåt baserat på data ett år tidigare. Eftersom en försäkringskund endast kan förkomma en gång i data från ett år så kan inte beroendet mellan observationer som vi har diskuterat förkomma i den modelleringen.

Resultatet i prediktionsförmåga av att anpassa modellerna på data från ett år (jämförelse mellan Tabell 5 och 7) blir nästan likvärdigt eller något bättre. Detta skulle då kunna vara en antydan om att effekter som fanns inom försäkringskontrakten leder till skattningar av parametrarna som inte riktigt fångar upp effekterna från endast de förklarande variablerna, utan också försäkringsspecifika effekter. Och eftersom vi ser att prediktionsförmågan inte verkar försämras drastiskt av att använda data från endast ett år tillbaka så kan det för en framtida användning av dessa modeller och användning av parametrarna i logistisk regression vara lämpligare att använda data endast ett år tillbaka.

7.3 Obalanserat data

Vårt försäkringsdata är mycket obalanserat, vi har endast 1.9% ickeförnyelser i vårt data. Vi tillämpar därför en metod som kallas under sampling för att låta ickeförnyelserna få större vikt i vår modellering. I den metod vi använder för att välja vilken andel vi ska sätta att ickeförnyelserna ska utgöra i vårt data uppstår ett visst moment 22. Vi anpassar nämligen modellerna till ett data som är balanserat så att ickeförnyelserna utgör 20% av data. Vi använder sedan modellerna vi har kommit fram till, baserat på detta data, för att studera vilka andelar som verkar lämpligast att använda. Vi kommer fram till 20% för logistisk regression, och 30% för Random Forest, men väljer ändå att använda 20% för båda modellerna så att de anpassas på samma data. En annan metod här hade eventuellt kunnat leda till att en annan andel av ickeförnyelser i data varit det mest lämpliga och eventuellt kunnat förbättra prediktionen ytterligare.

Balanseringen av data vid anpassningen av modellerna ser inte ut att förbättra resultatet för den logistiska regressionsmodellen, men för Random Forest-modellen. Minskningen i antalet observationer vi jobbar med ger oss möjligheten att bygga flera träd med färre observationer i löven än vad ett obalanserat data kunde göra, eftersom tidsåtgången för algoritmen blir så stor när vi försöker anpassa en Random Forest-modell med över 300 träd på obalanserat data. I Tabell 6 kan vi se att precisionen för att prediktera förnyelse $Y = 1$ minskar något när vi använder balanserat data, men vinsten blir en bättre prediktion för ickeförnyelserna, vilket var vad vi ville åstadkomma. Även om logistisk regression inte förbättrades något

av balanseringen i vårt fall så verkar den inte heller ha försämrats, och vinsten för oss blev tidsåtgången vid modelleringen.

7.4 Förklarande variabler

Ett av syftena med detta arbete var också att studera vilka förklarande variabler de två metoderna visar har ett starkt samband med sannolikhet för förnyelse och om valet av förklarande variabler skiljer sig mycket mellan de två metoderna.

Inom logistisk regression valde vi att konstruera univariata modeller för varje förklarande variabler med förnyelse som responsvariabel, för att sedan inkludera de som fick ett signifikant värde vid ett likelihoodkvottest i en större logistisk regressionsmodellen. Vi utför sedan en form av backwards selection där vi exkluderar insignifikanta variabler med utgång from signifikansnivån 5%. I båda stegen studera som kontinuerliga variabler istället ska inkluderas som kategoriska och delas upp på intervall. I Random Forest så inkluderas alla förklarande variabler på en gång, utan analys om de kontinuerliga ska inkluderas som kategoriska istället.

De variabler som exkluderas i logistisk regressionsmodellen är de gällande antalet skador som rapporterats och medelskadekostnaden inom försäkringskontraktet, nämligen variablerna 4-7. I Figur 7 kan vi även se att dessa inte rangordnas som variabler som påverkar Gini-index eller prediktionsmåttet mycket. Random Forest-modellen visar att bl.a. variablerna 15, 11 och 3 alla har en stor inverkan i modellens homogenitet och prediktionsförmåga, vilka vi alla har valt att inkludera i den logistiska regressionsmodellen och baserat på absolutvärdet av t-statistikan också värderar högt i modellen, se Figur 5. Några av de variabler som i båda modellerna visar sig ha en inverkan på förnyelsegraden är ålder, försäkringsduration, vilket bolag försäkringen finns inom och hur många försäkringar kunden innehar samtidigt inom något bolag i Dina-federationen.

En variabel som vi var speciellt intresserade av att studera var prisförändringen, eftersom vi i början av detta arbete resonerade kring om försäkringskunder med boendeförsäkringar är priskänsliga. Denna variabel blir signifikant i den logistiska regressionsmodellen när vi inkluderar den som kategorisk och Random Forest-modellen klassar den som en av de viktigare för homogenitet i löven och förbättrad prediktionsförmåga. Den är dock inte bland de med störst inverkan i den logistiska regressionsmodellen, utan variablerna ålder och betalningsmetod har t.ex. större inverkan. I den komplexa Random Forest-modell vi har konstruerat kan vi dock inte se exakt vilket samband som finns mellan prisförändringen och förnyelsegraden utan vidare analys av de predikterade förnyelsegrader som modellen ger oss. I den

logistiska regressionsmodellen kan vi se skattningen av parametrarna för de olika faktorerna tillhörande de intervall vi delade in prisförändringen i och se att oddset för förnyelse ökar något när prisförändringen är låg. Från båda modellerna kan vi då se en antydan till att kunder med boendeförsäkringar faktiskt skulle vara priskänsliga.

8 Slutsats

Anpassningen av en logistisk regressionsmodell och en Random Forest-modell visar båda att det finns ett samband mellan försäkrings- och kundspecifika egenskaper och sannolikheten att förnya sitt boendeförsäkringskontrakt vid försäkringsperioden slut. Några av de variabler som visar sig ha en inverkan är kundens ålder, försäkringsduration och hur många försäkringar kunden har samtidigt hos bolaget. Även prisförändringen visar sig ha en inverkan på sannolikheten för förnyelse av försäkringskontraktet i båda modellerna, vilket indikerar att det finns en priskänslighet hos boendeförsäkringskunder.

Eftersom vi har ett mycket obalanserat data där ickeförnyelserna endast utgör 1.9% så använde vi oss av balanserat data vid anpassningen av modellerna för att låta ickeförnyelserna i data få större vikt bland observationerna. Detta gjorde inte någon skillnad för den logistiska regressionsmodellen, men gav en bättre prediktionsförmåga för ickeförnyelserna inom Random Forest-modellen.

Modellernas prediktionsförmåga är något bättre för den logistiska regressionsmodellen om vi studerar förmågan att prediktera ickeförnyelser när vi har använt data från 2004-2015 för att prediktera förnyelsen för försäkringskontrakt under 2016. Random Forest modellen är istället något bättre på att prediktera förnyelserna. Om vi använder data från endast ett år tillbaka för att prediktera ett år framåt så blir modellerna mer likvärdiga i sin prediktionsförmåga och något bättre än tidigare anpassning på större data.

De två modellerna ger oss liknande resultat gällande vilka variabler som har en inverkan på förnyelsegraden, men olikheterna i uppbyggnaden av modellerna innebär för- och nackdelar med båda. För den logistiska regressionsmodellen måste vi kontrollera så att antaganden om fördelning, oberoende observationer och oberoende förklarande variabler håller, vilket vi inte behöver för Random Forest-modellen. Random Forest-modellen blir mycket omfattande vid stora datamängder och vi kan inte gör en lätt tolkning av vilka sambanden mellan förklarande variabler och responsvariabler är, vilket vi kan göra i den logistiska regressionsmodellen genom skattade parametrar och logaritmerade odds.

Dessa modeller skulle eventuellt kunna användas i kombination, genom att använda Random Forest som ett analysverktyg vid valet av förklarande variabler. Variablerna kan sedan inkluderas i en GLM där vi kan skatta parametrar som vi lätt kan tolka och använda för att beskriva sambanden mellan förklarande variabler och responsvariabler.

9 Vidare utveckling av arbetet

Som vi nämnt tidigare i arbetet så skulle överlevnadsanalys kunna tillämpas på samma frågeställning som detta arbete belyser. Då skulle vi istället låta försäkringsdurationen vara en tidsindikator så att vi följer ett försäkringskontrakt under en längre tid och även inkluderar annullation som ett möjligt utfall för kontraktet. Denna typ av modellering skulle då även kunna besvara när under året ett försäkringskontrakt kommer att avslutas, inte bara om det kommer att avslutas vid försäkringsperiodens slut som det har studerats i det här arbetet.

Det skulle även vara intressant att studera priselasticiteten närmare för boendeförsäkringarna, nu när vi ser att prisförändringen har en inverkan på om en boendeförsäkringskund väljer att förnya sitt kontrakt. Eventuellt skulle vi då behöva studera försäkringarna i större detalj än vad som har gjorts i detta arbete, eftersom ett priskänslighetsmått ges för olika prisnivåer och priserna kan skilja sig mellan villor, lägenheter och fritidshus. Ett priskänslighetsmått skulle då kunna hjälpa försäkringsbolag med sin prissättningsstrategi, då man kan se hur pass mycket en prisförändring kommer att påverka förnyelsen i olika grupper.

Arbetet kan även utvecklas för andra försäkringstyper. Då specifika kontrakt inom företagsförsäkringarna kan utgöra en större del av beståndet än vad en enskild boendeförsäkring gör så skulle det vara intressant att studera om branschtillhörighet har en inverkan på förnyelsegraden, för att då kunna anpassa strategin gällande företagsaffären. Eftersom det kan förekomma en mer aktiv upphandling av försäkringskontrakten gällande större företag så skulle det också vara intressant att studera om det förekommer en priskänslighet även för företagsförsäkringarna och hur den ser ut jämfört med boendeförsäkringarna.

Appendix

A Härledning av maximumlikelihoodekvationerna för logistisk regression

Denna härledning baseras på s. 137 i Agresti, 2013. Vi börjar med att skriva om sannolikhetsgenererande funktionen för en binomialfördelad variabel, Y, med parametrarna N och $\pi(x)$, i formen av en fördelning som tillhör den exponentiella familjen fördelningar som

$$\begin{aligned} P(Y = y) = p(y) &= \binom{N}{y}\pi(x)^y(1-\pi(x))^{N-y} \\ &= \exp\left\{\log\binom{N}{y} + y\log(\pi(x)) + (N-y)\log((1-\pi(x)))\right\} \\ &= \exp\left\{\log\binom{N}{y} + y\log\left(\frac{\pi(x)}{1-\pi(x)}\right) + N\log((1-\pi(x)))\right\}. \end{aligned}$$

Då fås att $\theta = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$, $a(\phi) = 1$ och $c(y,\theta) = \log\binom{N}{y}$ och vi vill veta $b(\theta)$, som fås genom

$$\begin{aligned} e^\theta = \frac{\pi(x)}{1-\pi(x)} &\implies e^\theta - e^\theta\pi(x) - \pi(x) + 1 = 1 \\ &\implies (1+e^\theta)(1-\pi(x)) = 1 \\ &\implies (1-\pi(x)) = \frac{1}{(1+e^\theta)} \\ &\implies \log(1-\pi(x)) = -\log(1+e^\theta), \end{aligned}$$

så att $b(\theta) = -N\log(1-\pi(x)) = N\log(1+e^\theta)$.

Då ses att $b'(\theta) = N\frac{e^\theta}{1+e^\theta}$ och om vi sätter in θ får vi $b'(\theta) = \mu = E[Y] = N\pi(x)$.

Andraderivatan av $b(\theta)$ ger oss variansfunktionen, som vi behöver för maximumlikelihoodekvationerna, som blir $b''(\theta) = v(\mu) = N\pi(x)(1-\pi(x))$. Vi behöver även

$$g'(\mu) = \frac{\delta}{\delta\mu}\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \frac{\delta}{\delta\mu}\log\left(\frac{\mu}{N-\mu}\right) = \frac{N}{N\mu(N-\mu)} = \frac{1}{N\pi(x)(1-\pi(x))}$$

Om vi sätter in allt i maximumlikelihoodekvationerna så fås

$$\sum_{i=1}^{N}\frac{y_i-\mu_i}{N\pi(x)(1-\pi(x))\frac{1}{N\pi(x)(1-\pi(x))}}x_{ij} = 0 \implies \sum_{i=1}^{N}(y_i-\mu_i)x_{ij} = 0.$$

B Skattade parametrar i logistisk regressionsmodell

Variabel		β_i	e^{β_i}	**95% konfidensintervall för** e^{β_i}
`intercept`		69.02	$9.46 \cdot 10^{29}$	$(4.6 \cdot 10^{25}, 1.95 \cdot 10^{34})$
`variabel1`		-0.03	0.97	(0.96,0.97)
`variabel2`	0	0	1	-
	1	0.13	1.14	(1.01,1.28)
`variabel3`	0	0	1	-
	1	0.98	2.67	(2.54,2.81)
`variabel9`	1	0.51	1.67	(1.56,1.80)
	2	0.37	1.44	(1.36,1.53)
	3	0.68	1.98	(1.85,2.12)
	4	0.10	1.10	(1.01,1.21)
	5	0.25	1.29	(1.22,1.36)
	6	0.62	1.85	(1.75,1.96)
	7	0	1	-
	8	0.64	1.90	(1.77,2.04)
	9	0.40	1.49	(1.38,1.60)
	10	0.59	1.81	(1.69,1.94)
	11	0.48	1.61	(1.51,1.71)
	12	0.46	1.58	(1.42,1.76)
`variabel10`	0	-0.30	0.74	(0.70,0.79)
	1	0.17	1.19	(1.14,1.24)
	2	0	1	-
`variabel11`		0.15	1.16	(1.16,1.17)
`variabel12`	0	0	1	-
	1	-0.15	0.86	(0.78,0.95)
`variabel13`		0.10	1.10	(1.09,1.11)
`variabel14`	0	0.26	1.30	(1.17,1.45)
	1	0.10	1.11	(1.04,1.18)
	2	0	1	-
`variabel15`		0.02	1.02	(1.022,1.024)
`variabel16`	0	0	1	-
	1	0.20	1.22	(1.18,1.27)
`variabel17`	0	-0.35	0.71	(0.69,0.75)
	1	0	1	-
`variabel18`	0	-0.09	0.91	(0.86,0.97)
	1	0	1	-

Tabell 8: *Skattade parametrar i logistisk regressionsmodell.*

Referenser

Agresti, Alan. 2012. *Categorical Data Analysis*. 3 edn. Hoboken, New Jersey: John Wiley and Sons, Inc.

Belsley, David A., Kuh, Edwin, & Welsch, Roy E. 1980. *Regression Diagnostics, Identifying Influential Data and Sources of Collinearity*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Breiman, Leo. 2001. *Random Forest*. `https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf`, URL-datum: 2017-01-25.

Dal Pozzolo, Andrea, Caelen, Olivier, A. Johnson, Reid, & Bontempi, Gianluca. 2015. Calibrating Probability with Undersampling for Unbalanced Classification. *2015 IEEE Symposium Series on Computational Intelligence*.

Dina Försäkringar. 2015. *Dina Försäkringar - Vår Historia*. 2015-03-15. `https://www.dina.se/om-oss/var-historia.html`, URL-datum: 2017-01-23.

Dina Försäkringar AB. 2016. *Årsredovisning 2016*. `https://www.dina.se/download/18.6d69adf115b874a570054d/1492694527737/Dina+F%C3%B6rs%C3%A4kring+AB+%C3%85rsredovisning_2016.pdf`, URL-datum: 2017-04-26.

Fahrmeir, Ludwig, Kneib, Thomas, Lang, Stefan, & Marx, Brian. 2013. *Regression*. Berlin: Springer.

Fox, John, & Monette, Georges. 1992. Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, **87**(417), 178–180.

Fu, Luyang, & Wang, Hongyuan. 2015. Estimating Insurance Attrition Using Survival Analysis. *Variance*, **8**(1), 55–72.

Hosmer, David W., Lemeshow, Stanley, & Sturdivant, Rodney X. 2013. *Applied Logistic Regression*. 3 edn. Hoboken, New Jersey: John Wiley and Sons, Inc.

James, Gareth, Witten, Daniela, Hastie, Trevor, & Tibshirani, Robert. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer. `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf`, URL-datum: 2017-01-27.

King, Gary, & Zeng, Langche. 2001. Logistic Regression in Rare Events Data. *Political Analysis*, **9**(2), 137–163.

Lindsey, James K. 1997. *Applying Generalized Linear Models*. 3 edn. New York: Springer.

Liu, Yang, Chawla, Nitesh V., & Harper, Mary P. 2006. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech and Language*, **20**, 468–494.

Ohlsson, Esbjörn, & Johansson, Björn. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*. 1 edn. Berlin: Springer.

Ruiz-Gazen, A., & Villa, N. 2007. Storms Prediction: Logistic Regression vs Random Forest for Unbalanced Data. *Case Studies in Business, Industry and Government Statistics*, **1**(2), 91–101.