Stockholms universitet

Undersökning om försäkringskunder med bank-engagemang är mindre riskbenägna än andra kunder i ett och samma bestånd

Anna Francesconi

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2020:9
Försäkringsmatematik
September 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Masteruppsats **2020:9**
http://www.math.su.se/matstat

Undersökning om försäkringskunder med bank-engagemang är mindre riskbenägna än andra kunder i ett och samma bestånd

Anna Francesconi*

September 2020

Sammanfattning

I denna studie har tre olika modeller tillämpats på ICA försäkrings databestånd för de tre olika grenarna inom motor-affären i syfte att undersöka om det finns skäl att tro att de utav försäkringstagarna som har ett engagemang även hos ICA banken inom samma koncern är mindre riskbenägna än andra kunder inom samma bestånd. De tre modellerna är Generaliserad linjär modell (GLM), Logistisk regression (ett specialfall av GLM) och Random forest (en trädbaserad machine-learning-algoritm). För att kunna besvara frågeställningen beräknades och jämfördes relationstalen för en binär variabel Bank, vars värde berodde på om en kund har engagemang i banken eller inte, i fallet med GLM och Logistisk regression och ett Variable Importance-mått i fallet med Random Forest. Datamängden delades upp i en träningsmängd, som alla modeller tillämpades på, och en testmängd vilken användes för att validera och jämföra modellerna sinsemellan. Det visade sig att för det givna databeståndet fungerade GLM bäst med slutsatsen att försäkringstagarna med bank-engagemang är mindre riskbenägna än andra kunder inom detta bestånd (med vissa reservationer).

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post:anna.francesconi@hotmail.com. Handledare: Filip Lindskog.

Abstract

In this thesis three different models were applied to the motorinsurance-population of ICA-försäkring. The aim is to explore the possibility amongst this population, the policyholders who are also involved with ICA-banken (belonging to the same corporate group) have a more risk-averse behaviour than the policyholders who are not involved with ICA-banken. The three models used are Generalized linear model (GLM), Logistic regression (special case of GLM) and Random Forest (a tree-based machine learning-algorithm). To answer our question, we computed and compared the relativities for a binary variable Bank – the value of which depends on whether or not the policyholder is involved with ICA banken – when using GLM and Logistic regression. In the case of Random Forest a Variable Importance-measure was used. The dataset was, for each of the three parts in motor-insurance, divided into two parts, one training-set on which the models were applied, and one testset which was used to validate and compare the models to each other. As it turned out, GLM was the model that worked best on the given dataset. It enabled us to conclude that the policyholders involved with ICA-banken have a more risk-averse behaviour than other policyholders in the same population (with reservations).

Förord

Detta är mitt examensarbete på masternivå (30 hp) inom Masterprogrammet i Försäkringsmatematik(Aktuarieprogrammet) vid matematiska institutionen på Stockholms Universitet.

Tack till Olov Dahlberg, Mattias Ling och Malin Zonabend på ICA försäkring för er stöttning för att hjälpa mig att komma igång.

Jag skulle vilja rikta ett tack till min handledare Filip Lindskog på universitetet för värdefull feedback och rådgivning under arbetets gång.

Till sist vill jag också tacka min man, Erik, med dig i mitt liv klarar jag allt.

Innehåll

1 Inledning **4**
1.1 Syfte 4
1.2 ICA försäkring 4
1.3 Sekretess 5

2 Teori **6**
2.1 Begrepp inom försäkring 6
2.1.1 Premieargument 6
2.1.2 Duration 6
2.1.3 Skada 6
2.1.4 Skadefrekvens 6
2.2 Motor-försäkring 6
2.3 Generaliserad Linjär modell (GLM) 7
2.3.1 Den slumpmässiga komponenten: Exponentiella spridningsmodeller (EDM) 9
2.3.2 Den systematiska komponenten: Länkfunktionen 9
2.3.3 Variabelsignifikans 10
2.3.4 Variabeltyper 11
2.3.5 Offsets 11
2.3.6 Sannolikhetsfördelningen av skadefrekvensen och överspridning 11
2.3.7 Parameterskattning, Log-likelihood och devians 12
2.3.8 Hypotestestning och skattning av ϕ 13
2.3.9 Korrelation bland prediktorer och Multikolinjäritet 14
2.3.10 Logistisk regression 15
2.4 Träd och ensemble-metoder 16
2.4.1 PCA 16
2.4.2 Binära Regressions-träd 17
2.4.3 Random forest 17
2.5 Modell-validering 20
2.5.1 GLM, devians-residualer och randomized quantile residuals 20
2.5.2 Logistisk regression och ROC-kurvor 22
2.5.3 Random Forest och Out-of-Bag 23
2.5.4 Validation set approach och Actual vs. Predicted 24

3 Data **25**
3.1 Försäkringsdata 25
3.2 Uppdelning av data 26
3.3 Oberoende observationer 26
3.4 Variabler 26
3.5 Dataanalys 27

4 Metod och resultat **29**
4.1 Multiplikativ GLM-modell 29
4.2 Logistisk regression 34
4.3 Random Forest 34
4.3.1 Implementation i R 34
4.3.2 Metod 34

5 Validering och jämförelse **38**
5.1 Validering av Multiplikativ GLM 38
5.2 Validering av Logistisk regression 39
5.3 Validering av Random Forest 40
5.4 Jämförelse av alla tre modellerna 41

6 Diskussion och slutsats **44**
6.1 Kritisk diskussion och felkällor 44
6.2 Slutsats 45
6.3 Vidare utveckling av arbetet 45

7 Appendix **46**

8 Källhänvisningar **56**

1. Inledning

Försäkringsbolag förmedlar säkerhet till samhället genom att erbjuda skydd mot ekonomiska förluster. De tillåter individer att byta osäkerhet mot säkerhet, genom att överföra risken på försäkringsbolaget i utbyte mot en premie. Ett försäkringsbolag sätter priset för ett försäkringskontrakt innan dess faktiska kostnad är känd. På grund av detta är det av stor vikt att försäkringsbolaget ordentligt bedömer riskerna i dess portfölj. Försäkringsbranschen är mycket data-driven och förlitar sig på dator-algoritmer för beslutsfattande. För att kunna prissätta ett kontrakt predikterar man kostnaden av förlusten y för varje försäkringstagare baserat på hens observerbara egenskaper $\boldsymbol{x}$. Därför utvecklar försäkringsbolaget en prediktiv modell f, som mappar riskfaktorerna $\boldsymbol{x}$ mot den predikterade förlusten $\hat{y}$ genom att sätta $\hat{y} = f(\mathbf{x})$. Generaliserade linjära modeller (GLMs), introducerade av Nelder och Wedderburn (1992), är industri-standard för att utveckla analytiska prissättningsmodeller. Men teknologisk framfart har ökat populariteten för Machine learning och Big data-analys, och därmed har landskapet för prediktiv modellering förändrats.

1.1 Syfte

Jag gör mitt arbete på ICA försäkring och har fått alla data för deras motoraffär att jobba med. Det finns indikationer i dessa data på att försäkringskunder som även är kunder hos ICA-banken (i samma koncern) är mindre riskbenägna (dvs. så kallade 'bra' kunder) än övriga kunder i detta bestånd. Syftet är att undersöka detta mer på djupet och se om jag kan underbygga detta påstående ytterligare. Jag kommer att använda mig av flera olika modeller för detta: GLM - multiplikativ modell, Logistisk regression och en Machine learning-modell (Random Forest). Om så är fallet så är det intressant att göra en prisbedömning och ev. ge rabatterat pris till kunder med bankengagemang för att på så vis locka till sig fler av samma kundtyp.

1.2 ICA försäkring

ICA Försäkring började sälja försäkringar den 1 oktober 2015. Försäkringsbolaget, med säte i Solna, är dotterbolag till ICA Banken och ingår i ICA Gruppen. ICA försäkring har växt snabbt och har i skrivande stund ca 160 000 kunder [16].

1.3 Sekretess

Eftersom detta arbete har gjorts på ett företag så kommer alla variabler utom skadefrekvensen, bank-variabeln och egna variabler att pseudonymiseras.

2. Teori

2.1 Begrepp inom försäkring

2.1.1 Premieargument

För varje försäkringskontrakt bestäms premien av värdena på ett antal variabler, *premieargumenten*. För att skatta detta förhållande tillämpas en statistisk modell (t.ex. GLM). Traditionellt bildas premieargument genom att dela in variablerna i klasser, om de inte redan är kategoriska. Anledningen till detta är delvis att vi sällan har en linjär relation och delvis regeln enligt vilken premien beräknas för en försäkringstagare, *tariffen*. Försäkringskontrakt som tillhör samma klass för varje premieargument sägs tillhöra samma *tariffcell* och ges samma premie.

2.1.2 Duration

Den tid som ett försäkringskontrakt gäller, vanligen mätt i år.

2.1.3 Skada

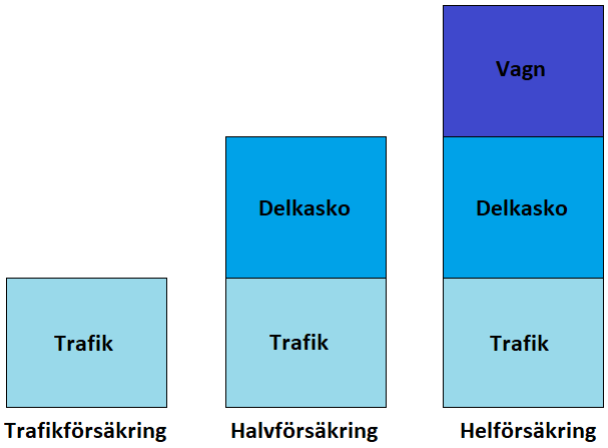
En skada är en händelse rapporterad av försäkringstagaren, för vilken hen kräver ekonomisk kompensation.

2.1.4 Skadefrekvens

Antalet skador dividerat med durationen, dvs. det genomsnittliga antalet skador per tidsperiod. Denna storhet är högt korrelerad med den risk-aversivitet vi vill undersöka.

2.2 Motor-försäkring

Motorförsäkringen består av tre delar: Trafik-,Delkasko- och Vagnskadeförsäkring. Traditionellt så tecknar försäkringstagarna sedan trafikförsäkring, halvförsäkring (delkasko, men även trafik ingår) eller helförsäkring (alla tre delar ingår).



Figur 1. *Illustration av de olika typerna av motorförsäkring.*

Trafikförsäkring är ett krav enligt lag för alla bilar och är ett skydd för dina medtrafikanter i en eventuell olycka. När du köper en bil behöver du ha trafikförsäkring redan från första dagen. Trafikförsäkringen är alltså det lägsta försäkringsskydd man kan välja till sin bil.
Delkasko-försäkringen hos ICA försäkring täcker till exempel kostnader för oförutsedda fel på motordelar, växellåda och klimatanläggning under de första åtta åren, eller 10 000 milen. Även allriskskydd, bärgning och hemtransport ingår i försäkringen, liksom reparation eller byte av glasruta [15].
Till skillnad från helförsäkringen saknar halvförsäkringen vagnskadeskydd. Det betyder att halvförsäkringen inte ger ersättning för skador på försäkringstagarens bil vid trafikolycka, skadegörelse eller olyckshändelse. Köper man en fabriksny bil ingår oftast vagnskadegaranti, ibland kallad personvagnsgaranti, de tre första åren. Under den tiden ger en halvförsäkring ett fullgott skydd. Efter det rekommenderas det att välja helförsäkring fram till att bilen är minst tio år gammal, eller när värdet blivit så lågt att det inte lönar sig med vagnskadeskydd.

2.3 Generaliserad Linjär modell (GLM)

Generaliserade linjära modeller (GLMs) är ett medel för att kunna modellera förhållandet mellan en variabel vars utfall vi önskar prediktera (responsvariabeln) och en eller flera prediktorer (premieargument). Den predikterade variabeln kallas *responsvariabel* och betecknas y medan prediktorerna betecknas med $\boldsymbol{x}$. För kvantitativa responsvariabler producerar en GLM-modell en skattning av utfallets väntevärde. För kvalitativa variabler kan en GLM tillämpas för att skatta sannolikheten att en händelse inträffar (ett exempel på detta är logistisk regression).
I en GLM antas utfallet från responsvariabeln drivas av en *systematisk* komponent såväl som en *slumpmässig* komponent.
Den systematiska komponenten svarar mot den andel av variationen i utfallen som är relaterade till värdena på prediktorerna. Till exempel kanske vi tror att förarens ålder påverkar den förväntade skadefrekvensen för en individuell motorförsäkring. Om förarens ålder inkluderas som prediktor i frekvensmodellen, så

är den effekten del av den systematiska komponenten. Den slumpmässiga komponenten är andelen av utfallet som drivs av andra orsaker än prediktorerna i modellen. Detta inkluderar omständigheter som är oförutsägbara ens i teorin.

GLM är en rik klass av statistiska metoder, vilken generaliserar de ordinära linjära modellerna i två riktningar:
- *Sannolikhetsfördelning.* Istället för att anta normalfördelningen, jobbar GLMs med en generell klass av fördelningar (EDM), vilken innehåller ett antal diskreta och kontinuerliga fördelningar som specialfall, bland annat normal-, Poisson- och gamma-fördelningarna.
- *Modell för väntevärdet.* I linjära modeller är väntevärdet en linjär funktion av prediktorerna $\boldsymbol{x}$. I en GLM är någon monoton transformation av väntevärdet en linjär funktion av $\boldsymbol{x}$, med de linjära och multiplikativa modellerna som specialfall.

Statistisk modellering, inkl. GLM, gör vanligtvis antaganden om den slumpmässiga process som genererar data. Till exempel kan det antas att logaritmen av en variabel är approximativt normalfördelad. Fördelningsantaganden kontrolleras genom att jämföra empiriska percentil-ranger mot de som beräknas grundat på den antagna fördelningen. (s. 12 [10])

De grundläggande antaganden som görs som förberedelse för GLM är
(1) Oberoende mellan olika försäkringskontrakt.
(2) Oberoende mellan tidsperioder (försäkringskontraktens durationer).
(3) Homogenitet. Försäkringskontrakt inom samma tariffcell har samma fördelningsfunktion.

Begränsningar som detta medför:
(1) En GLM tilldelar data full trovärdighet (kredibilitet)
Skattningarna som produceras av en GLM anpassas under antagandet att data har full trovärdighet för varje parameter. För någon kategorisk variabel i modellen så är skattningen av koefficienten för varje nivå den som bäst passar träningsdata, där ingen hänsyn har tagits till glesheten av data som den baseras på. Exempelvis så ges en tariffcell med väldigt få observationer lika stor trovärdighet som en cell med många observationer.
Exempel: Antag att vi använder en GLM för att skatta medelskadan, och GLM inkluderar bara en prediktor: en kategorisk variabel med fem nivåer A-E. Datavolymen varierar mycket per nivå, och den nivå som har minst volym har bara 8 skador. Efter att ha kört denna modell kommer prediktionen för varje risk att vara det totala genomsnittet av medelskada. Den indikerade relativiteten för E baseras enbart på den genomsnittliga medelskadan för dessa 8 observationer. Standardfel och p-värde blir stora, och på detta sätt 'varnar' GLM för att skattningen inte är helt trovärdig - men gör inget åt det.

(2) En GLM antar att slumpmässigheten hos utfallen är okorrelerade
Ett annat viktigt antagande inbyggt i GLMs är att den slumpmässiga komponenten av utfallet av responsvariabeln är okorrelerad bland observationerna i träningsmängden. Notera benämningen 'slumpmässig komponent' i den meningen - med det menas andelen av utfallet som drivs av orsaker som inte finns i vår modell.

Detta antagande motsägs om det existerar grupper av observationer som troligen kommer att ha liknande utfall, möjligen på grund av någon dold variabel som inte fångas upp av vår modell.
Ett exempel på när detta kan uppstå i försäkringsmodeller:
Datamängden i en GLM kommer att omfatta flera år av försäkringsdata. Alltså kommer det att finnas många fall där distinkta observationer egentligen är multipla förnyelser av samma kontrakt. Dessa observationer kommer troligen att ha korrelerade utfall [5].

Som tidigare nämnt brukar GLM vara ganska robust och klarar av att hantera smärre överträdelser av dess antaganden. Men det är viktigt att komma ihåg att parameter-skattningar och signifikans-statistikor för en GLM alla härleds som om alla de slumpmässiga utfallen var oberoende.

2.3.1 Den slumpmässiga komponenten: Exponentiella spridningsmodeller (EDM)

I en GLM modelleras y - responsvariabeln - som en slumpvariabel som följer en sannolikhetsfördelning. Denna fördelning antas vara medlem av den exponentiella familjen av fördelningar (EDM). EDM är en klass av fördelningar som har särskilda egenskaper som är användbara när man anpassar GLMs. EDM generaliserar normalfördelningen som används i linjära modeller. Sannolikhetsfördelningen av en EDM ges av följande frekvensfunktion,

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\Big(\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i,\phi,\omega_i)\Big). \tag{2.1}$$

Här är θ_i en parameter som tillåts att bero på i, medan spridningsparametern $\phi > 0$ är densamma för alla i. Den så kallade *kumulantfunktionen* $b(\theta_i)$ antas två gånger kontinuerligt differentierbar, med inverterbar förstaderivata. För varje val av en sådan funktion, får vi en familj av sannolikhetsfördelningar, t.ex. normal-, Poisson- och gamma-fördelningar. Givet valet av $b(\theta_i)$ är fördelningen helt specificerad med parametrarna θ_i och ϕ. Funktionen $c(\cdot,\cdot,\cdot)$, vilken ej beror på θ_i, är inte av något större intresse inom GLM-teori.

2.3.2 Den systematiska komponenten: Länkfunktionen

En GLM modellerar förhållandet mellan μ_i (modellprediktionen, ej att förväxla med y under förra rubriken) och prediktorerna som följer:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \tag{2.2}$$

Denna ekvation säger att någon specificerad transformation av μ_i (betecknad $g(\mu_i)$) är ekvivalent med interceptet (betecknat β_0) plus en linjärkombination av prediktorerna och koefficienterna, vilka betecknas $\beta_1, ..., \beta_p$.
Transformationen av μ_i representerad av funktionen $g(\cdot)$ i vänsterled i ekvationen ovan kallas för *länk-funktion*.
Länkfunktionen $g(\cdot)$ tjänar alltså för att ge flexibilitet när man relaterar modellprediktionen med prediktorerna: snarare än att kräva att väntevärdet av responsvariabeln ska vara direkt ekvivalent med en linjärkombination av prediktorerna (som i fallet med vanlig linjär modell).

2.3.3 Variabelsignifikans

En viktig fråga för varje prediktor är om skattningen av koefficienterna är rimligt nära de 'sanna' koefficienterna? Och, kanske ännu viktigare, har prediktorn någon effekt på utfallet?
Vi kan använda oss av flera typer av statistikor för att få hjälp med att besvara dessa frågor, bl.a. standardfelet, p-värde och konfidensintervall.

Standardfel
De skattade koefficienterna är resultatet av en slumpprocess. Standardfelet är den skattade standardavvikelsen av den slumpprocessen. Till exempel, ett standardfel på 0.15 som tilldelas en koefficient-skattning kan tänkas på som följer: om denna process upprepades många gånger, så skulle standardavvikelsen för de resulterande skattningarna för koefficienten vara ungefär 0.15.
Ett litet standardfel indikerar att den skattade koefficienten förväntas vara nära den sanna koefficienten. Å andra sidan, ett stort standardfel säger oss att ett stort antal skattningar kan erhållas genom slumpmässighet, och gör det med andra ord mindre troligt att skattningen vi har är nära det sanna värdet.
Standardfelet är också relaterat till det skattade värdet av ϕ (spridningsparametern): ju större skattning av ϕ, ju större kommer standardfelen att vara. Detta beror på att ett större ϕ implicerar mer varians i slumpmässigheten av utfallen, vilket skapar mer 'brus' och resulterar i större standardfel.

p-värde
För en given koefficient-skattning så är p-värdet en skattning av sannolikheten att värdet av den magnituden uppstår av ren slump. Till exempel, antag att en viss variabel i vår modell ger en koefficient på 1.5 med p-värdet 0.0012. Detta indikerar att om denna variabels sanna koefficient är noll, så är sannolikheten 0.0012 att få en koefficient 1.5 eller högre av ren slump. I detta fall är det rimligt att dra slutsatsen att eftersom chansen för ett sådant värde är liten är det troligt att resultatet speglar en verklig underliggande effekt - dvs. den sanna koefficienten är inte noll. En sådan variabel sägs vara signifikant.

Konfidensintervall
Ibland är en naturlig fråga: vid vilket *intervall* av variabler, om hypotetiserade, skulle inte förkastas vid vår valda p-värdes-tröskel? Detta intervall kallas *konfidensintervall*, och kan ses som ett rimligt intervall av skattningar för koefficienten.

Konfidensintervall för GLM parameter-skattningar kan härledas från vanlig maximum likelihood-teori, enligt vilken maximum-likelihood-skattningarna (ML-skattningarna) under allmänna villkor är asymptotiskt normalfördelade och väntevärdesriktiga, med kovariansmatrisen ekvivalent med inversen av Fisher informationen $\boldsymbol{I}$.
I allmänhet definieras Fisher informationen som den negativa andraderivatan av log-likelihood-funktionen

$$I(\theta) = -\frac{d^2 l(\theta)}{d\theta^2}. \text{ [6]}$$

Detta asymptotiska resultat motiverar approximationen $\hat{\boldsymbol{\beta}} \overset{\text{d}}{\approx} N(\boldsymbol{\beta}; \boldsymbol{I}^{-1})$, där $\overset{\text{d}}{\approx}$ betecknar approximativ fördelning. Enligt s.46 i [14] kan vi beräkna konfidensin-

tervall för (pris-) relativiteterna (mäter den relativa skillnaden i förhållande till bascellen) $\gamma_i = \exp(\beta_j)$ som följer:
Låt $\{c_{j,k}\}$ beteckna elementen i kovariansmatrisen $\boldsymbol{C} = \boldsymbol{I}^{-1}$. Ett approximativt 95%-konfidensintervall (a, b) för β_j är då $\hat{\beta}_j \pm 1.96\sqrt{c_{jj}}$.
Ett approximativt 95%-konfidensintervall för γ_j ges av $(\exp(a), \exp(b))$. Notera att dessa gränser inte är symmetriska runt skattningen $\hat{\gamma}_j$. I fallet med logistisk regression ges konfidensintervallet för γ_j istället av $(\text{invlogit}(a), \text{invlogit}(b))$ (se Teori 2.3.10).

2.3.4 Variabeltyper

Förklaringsvariabler i en GLM klassificeras som antingen kontinuerliga eller kategoriska, och var och en av dessa typer behandlas olika.
En *kontinuerlig* variabel är en numerisk variabel som representerar ett mått på en kontinuerlig skala.
En *kategorisk* variabel är en variabel som antar ett av två eller flera möjliga värden, och därmed tilldelar varje risk till en kategori. Dessa kategorier brukar kallas nivåer (eller klasser).

Behandling av kontinuerliga variabler
När en log-länk används (som i fallet med en multiplikativ modell) är det nästan alltid lämpligt att ta den naturliga logaritmen av kontinuerliga förklaringsvariabler innan dessa inkluderas i modellen, snarare än att ha med dem i sin ursprungliga form i modellen. Detta gör att skalningen blir rätt.

Behandling av kategoriska variabler
En av nivåerna sätts som *bascell*. Man ersätter kolumnen i input-datamängden som innehåller den kategoriska variabeln med en serie av indikator-kolumner, en för varje nivå av den variabeln förutom bascellen. Var och en av dessa kolumner antar värdena 0 eller 1, där 1 indikerar medlemskap på den nivån. Dessa kolumner behandlas som separata prediktorer, och var och en får sin egen koefficient i output. Denna resulterande datamängd kallas *designmatrisen*.

2.3.5 Offsets

Exponering är ett mått på hur man vill dela upp sina data. Offset är en modelleringsteknik i Poissonregression som tillåter att man modellerar för kvoter utan att behöva skapa en ny statistisk struktur. Poissonregression används typiskt för att modellera räknedata (som skade-antal). Men ibland är det mer relevant att modellera kvoter istället. Detta är relevant när exempelvis individer inte följs under lika lång tid. Så istället för att ha $\ln(\mu) = \beta_0 + \beta_1 x$, har man istället

$$\ln(\frac{\mu}{t}) = \beta_0' + \beta_1' x \quad \Longleftrightarrow \quad \ln(\mu) = \beta_0' + \beta_1' x + \ln(t). \tag{2.3}$$

2.3.6 Sannolikhetsfördelningen av skadefrekvensen och överspridning

Låt $N(t)$ vara antalet skador för ett individuellt försäkringskontrakt under tidsintervallet $[0, t]$, med $N(0) = 0$. Den stokastiska processen $\{N(t); t \geq 0\}$

kallas för *skadeprocessen*. Enligt s.18 [14], har man visat att under antaganden som är nära våra grundläggande modellantaganden om homogenitet och tids-oberoende (plus ett antagande om att skador inte "klustrar", dvs. är ihopklumpade), är skadeprocessen en Poisson process. Detta motiverar oss att anta en Poissonfördelning för antalet skador från ett individuellt försäkringskontrakt under en given tidsperiod. Vi är dock mer intresserade av fördelningen för skadefrekvensen $Y_i = X_i/\omega_i$, och eftersom denna fördelning är en transformation av Poissonfördelningen brukar den kallas för den *relativa Poissonfördelningen*.

Man finner ofta att just skadefrekvensen har en varians som är större än väntevärdet, ett fenomen som kallas *överspridning*. Överspridning uppstår huvudsakligen för att det, förutom den naturliga variansen från Poissonprocessen, finns en annan källa till varians: variationen i risknivå bland försäkringstagarna själva. I statistiska termer: förutom Poisson-variansen finns det en varians i Poissonväntevärdet μ bland risker.
Därför använder vi som nämnt den relativa Poisson-fördelningen, annars är det mycket troligt att variansen kommer att underskattas och därmed snedvrida modell-diagnostiska mått som standardfel och p-värde.

Eftersom den empiriska variansen för alla tre grenarna var mycket större än det motsvarande empiriska väntevärdet för varje gren, så kan vi anta att vi här har att göra med överspridning.
Ett vanligt sätt att modellera överspridning är att se väntevärdet, Λ i Poissonprocessen som en slumpvariabel. För att specificera modellen fullständigt, måste en fördelning för Λ ansättas, den så kallade *blandningsfördelningen*. Den i särklass mest kända modellen får man då man antar att Λ har en gammafördelning. Då har vi en så kallad *negativ binomialfördelning*, som allmänt definieras på följande sätt [9]:
En diskret slumpvariabel X sägs vara negativt binomial-fördelad med parametrar $\alpha > 0$ och $\beta > 0$ om dess täthetsfunktionen ges av

$$p_k(\alpha, \beta) = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{k}.$$

Den negativa binomialfördelningen har två parametrar, vilket ger oss möjligheten att få de två första momenten att överensstämma med den empiriska fördelningen. Momentmetoden ger

$$\begin{cases} \hat{\alpha} = \bar{x}^2/(s^2 - \bar{x}), \\ \hat{\beta} = \bar{x}/(s^2 - \bar{x}). \end{cases}$$

2.3.7 Parameterskattning, Log-likelihood och devians

GLM-mjukvara ger ett antal statistiska mått på hur väl modellen passar data. De viktigaste sådana måtten är *log-likelihood* och *devians*.

För någon given mängd koefficienter, så implicerar en GLM ett probabilistiskt väntevärde för varje observation. Det, tillsammans med spridningsparametern och vald fördelningsform, implicerar en sannolikhetsfördelning. Det är därför möjligt att beräkna sannolikheten (eller tätheten) som GLM-modellen skulle tilldela förekomsten av ett visst utfall som faktiskt har inträffat. Multiplicerar

vi dessa värden över alla observationer så får vi sannolikheten för alla historiska utfall som förekommer; detta värde kallas *likelihood*.
En GLM anpassas genom att hitta den mängd parametrar för vilken likelihooden har störst värde. Eftersom likelihood vanligtvis är ett extremt litet tal, så använder man oftast log-likelihood istället för att underlätta beräkningarna.

Extremfallet i den lägre änden av skalan är null-modellen, eller en hypotetisk modell utan prediktorer - endast intercept. En sådan modell producerar samma skattning för alla observationer: det totala väntevärdet (the grand mean").
Extremfallet åt det andra hållet ligger i den *mättade* modellen, eller en hypotetisk modell med ett ekvivalent antal prediktorer som observationer i datamängden. En sådan modell ger ett ekvationssystem med n ekvationer och n okända, och därför en perfekt lösning. Den skulle troligen vara oanvändbar, pga. extrem överanpassning så är det egentligen bara ett komplicerat sätt att skriva upp historiska data.

Deviansen, D, för en GLM definieras som följer:

$$D = 2 \cdot (l_{\text{mättad}} - l_{\text{modell}}),$$

där $l_{\text{mättad}}$ är log-likelihooden av den mättade modellen, och l_{modell} är log-likelihooden av modellen som värderas. Det vill säga deviansen är 'distansen' mellan den mättade modellen och den faktiska modellen. De anpassade GLM koefficienterna är de som minimerar deviansen. Tidigare sa vi att GLM anpassas genom att maximera log-likelihood, och faktum är att dessa två påståenden är ekvivalenta: maximera log-likelihood och minimera deviansen.
Deviansen för den mättade modellen är noll, medan deviansen för null-modellen kan ses som den totala deviansen inuti data.

2.3.8 Hypotestestning och skattning av ϕ

Statistiska test hanterar problemet med att ta beslut. Antag att vi har en observation x av en slumpvektor $X \sim F_\theta$ med en given men okänd parameter θ som ligger i en given mängd Θ av möjliga parametrar. Målet är att testa huruvida parametern θ som har genererat x kan tillhöra någon delmängd $\Theta_0 \subset \Theta$.
Antag att vi vill kolla huruvida x kan ha genererats av en given parameter θ_0.

Nollhypotes H_0: $\theta = \theta_0$.
Tvåsidig alternativ hypotes H_1: $\theta \neq \theta_0$.

Då bygger vi sedan en test-statistika $T(X)$ vars fördelningsfunktion är känd under nollhypotesen H_0 och vi betraktar frågan huruvida $T(x)$ antar ett icke troligt värde under nollhypotesen. Man väljer en signifikansnivå $q \in (0, 1)$ och för denna en kritisk region. Noll-hypotesen förkastas då om $T(x)$ faller inom denna region. I praktiken beräknar man ofta ett så kallat *p-värde* som tidigare nämnt. Detta betecknar den kritiska sannolikheten vid vilken nollhypotesen förkastas.

Likelihood ratio test (LRT)

Låt $l(\hat{\boldsymbol{\mu}})$ beteckna log-likelihooden för vår data som en funktion av den skattade väntevärdesvektorn $\hat{\boldsymbol{\mu}}$ (där $\hat{\boldsymbol{\mu}}$ som tidigare nämnt är modellprediktionerna). Den *skalade deviansen* D^* är ett goodness-of-fit-mått och definieras som likelihood-ratio-test (LRT)-statistikan av modellen av intresse. LRT-statistikan är två gånger logaritmen av likelihoodkvoten, dvs.

$$D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2[l(\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}})]. \tag{2.4}$$

Hierarkiskt test

Som en hjälp när man ska bestämma huruvida en prediktor ska inkluderas i en GLM-modell kan vi utföra ett hypotestest. Detta kan göras med ett Likelihood ratio-test av de två modellerna mot varandra, med och utan den specifika prediktorn. Modellerna måste vara hierarkiskt ordnade så att nollhypotesen specificerar en submodell av den ursprungliga modellen, genom att sätta vissa parametrar i den ursprungliga modellen till noll.

Betrakta två modeller H_s och H_r, sådana att $H_s \subset H_r$. Låt $\hat{\boldsymbol{\mu}}^{(r)}$ vara ML-skattningarna under H_r och liknande för H_s. Då är LRT-statistikan för att testa H_s mot H_r:

$$D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(s)}) - D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(r)}),$$

där D^* står för den skalade deviansen.

2.3.9 Korrelation bland prediktorer och Multikolinjäritet

Frekvent så uppvisar prediktorerna som ingår i en GLM korrelation bland dem. Om sådan korrelation är måttlig kan GLM hantera det, men om exempelvis korrelationen mellan två prediktioner är stor kan vi få problem med GLM-modellen eftersom den höga korrelationen betyder att mycket av samma information kommer in i modellen flera gånger. Sådan instabilitet i en modell bör undvikas.

För att upptäcka eventuell hög korrelation kan man t.ex. undersöka två-vägs korrelationstabeller eller scatterplots för att se om det finns ett mönster/trend. Om man då upptäcker hög korrelation i en grupp av prediktorer kan man exempelvis ta bort alla prediktorer i denna grupp utom en från modellen. Även om detta är den enklaste approachen så är en potentiell nackdel att det kan finnas unik information i de individuella prediktorer som inte tas med i modellen.

Enkel korrelation mellan par av prediktorer är förhållandevis enkelt att upptäcka genom att använda en korrelationsmatris. Ett mer subtilt potentiellt problem kan existera där två eller fler prediktorer i en modell är starkt prediktiva av en tredje, en situation känd som multikollinearitet. Samma instabilitetsproblem som redan nämnts kan bli resultatet eftersom informationen som den tredje variabeln innehåller också finns i modellen i form av *kombinationen* av de andra två variablerna. Men variabeln behöver inte vara högt korrelerad med någon av de andra två prediktorerna individuellt, och därför dyker inte denna effekt upp i en korrelationsmatris.

En användbar statistika för att upptäcka multikollinearitet i en mängd av oberoende variabler är VIF (variance inflation factor). VIF beräknas för varje oberoende variabel enligt:

$$VIF = \frac{1}{1 - R^2}, \tag{2.5}$$

där R^2 är det statistiska måttet på hur nära data är en anpassad kurva och definieras

$$R^2 = 1 - \frac{\sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\text{felkvadratsumman}}{\text{totalkvadratsumman}}. \tag{2.6}$$

GVIF är kvadratroten av VIF för individuella prediktorer och kan alltså användas ekvivalent. Mer generellt så består GVIF-faktorer av VIF justerade för frihetsgrader.

$$GVIF = VIF^{1/2*Df},$$

där Df står för antalet frihetsgrader (degrees of freedom). Variabler som har mer än en frihetsgrad evalueras med hjälp av GVIF. En vanlig statistisk tumregel är att ett VIF/GVIF-värde större än 10 är att betrakta som högt. Där höga VIF/GVIF-värden indikerar att det viktigt att titta närmare på kollinearitetstrukturen för att kunna göra ett informerat beslut om hur det bäst ska hanteras i modellen. [17]

2.3.10 Logistisk regression

För vissa modeller, så är respons-variabeln som vi önskar prediktera inte ett numeriskt värde, utan snarare skeende eller icke-skeende av en händelse. Sådana variabler kallas binära. För att modellera ett sådant scenario i en GLM, sätts fördelningen för responsvariabeln till binomialfördelningen. Väntevärdet av binomial fördelningen - dvs. prediktionen som genereras av modellen - är sannolikheten att händelsen kommer att inträffa.

Det finns flera länk-funktioner tillgängliga för detta syfte, men den vanligaste är *logit*, definierad som

$$g(\mu) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta x \quad \Rightarrow \pi = \frac{e^{x\beta}}{1 + e^{x\beta}}. \tag{2.7}$$

Logit kan tolkas som logaritmen av oddset, där oddset definieras som kvoten av sannolikheten för skeende mot sannolikheten för icke-skeende, eller $\frac{\pi}{1-\pi}$. Oddset är ett alternativt sätt att beskriva sannolikhet, vilken - till skillnad från sannolikhet som måste ligga i intervallet $[0, 1]$ - är obegränsad i positiv riktning.
Exponentierar vi bägge sidor av den fulla specifikationen av en logistisk regressionsmodell,

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \tag{2.8}$$

blir den logistiska GLM-ekvationen en multiplikativ serie av termer som producerar oddsen för skeende. Detta leder till en naturlig tolkning av koefficienterna

som att de beskriver effekten av förklaringsvariablerna på oddset.

Parametrar i en logistisk regression tolkas som följer. Betrakta en enda kontinuerlig förklaringsvariabel x och alltså

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x. \tag{2.9}$$

Exponentiering av bägge sidor visar att oddset är $e^{\alpha+\beta x}$. Antag att x ökas med en enhet, då förändras oddset från $e^{\alpha+\beta x}$ till $e^{\alpha+\beta(x+1)} = e^{\beta}e^{\alpha+\beta x}$. Med andra ord, att öka förklaringsvariabeln med en enhet multiplicerar oddset med e^{β}. Om β är litet så har vi att $(e^{\beta}-1) \approx \beta$ och i detta fall så är 100β den approximativa procentuella chansen i odds när förklaringsvariabeln ökar med en enhet. Om $\beta < 0$ så blir effekten av en ökning av x att oddset minskar för en händelses förekomst. Motsatsen då $\beta > 0$ är effekten av en ökning av x, en ökning av oddset. Slutligen, om $\beta = 0$ så är $e^{\beta} = 1$ och det blir ingen effekt på oddset.

När försäkringskontrakt är exponerade för risk t, där $0 < t \leq 1$, så är sannolikheten för en skada proportionellt reducerad med risktiden (den tid kontraktet var utsatt för risk). Responsfördelningen är $\text{Bin}(1, t\pi)$ och sannolikhetsfunktionen är:

$$f(y) = (t\pi)^y(1-t\pi)^{1-y}, \quad y = 0, 1. \tag{2.10}$$

En rimlig skattning av π är det genomsnittliga antalet skador viktade mot exponeringen.
Den lämpliga logistiska länken baseras på π, sannolikheten för en skada över hela perioden. Definierar $\pi^* = t\pi$:

$$\ln\frac{\pi^*/t}{1-\pi^*/t} = x\beta \quad \Rightarrow \pi^* = t\frac{e^{x\beta}}{1+e^{x\beta}}. \tag{2.11}$$

Justeringen för exponeringen uppnås därför genom en enkel modifiering av logitlänken.

2.4 Träd och ensemble-metoder

Här kommer vi till största delen att följa boken *'An Introduction to Statistical Learning'* [8].

2.4.1 PCA

Man kan för-processa data genom att använda dimensions-reduceringstekniker så som exempelvis PCA (principal component analysis), vilken i grova drag innebär att man simplifierar en datamängd genom att transformera originalvariablerna till ett mindre antal 'principal components'. Dessa i sin tur är den underliggande strukturen i data. De representerar de riktningar i vilka det finns mest varians, dvs. där data är mest utspridd. Detta betyder att vi försöker hitta den räta linje som bäst sprider ut data när den är projekterad på denna linje. Detta blir då den första principal komponenten. Denna linjära transformation

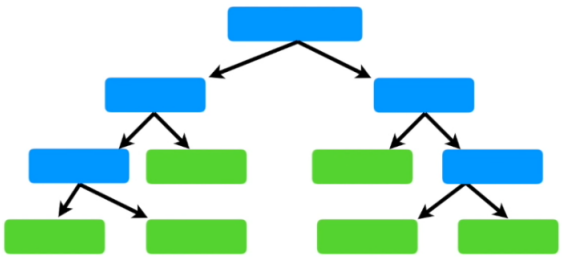
anpassas till detta dataset i ett nytt koordinatsystem på ett sådant sätt att den mest signifikanta variansen hittas på den första koordinaten, och varje efterföljande koordinat är ortogonal (vinkelrät) mot den förra och har lägre varians.

För att lättare se hur dessa principal components förhåller sig till varandra kan man använda sig av en scree-plot, som är en plot av principal komponenternas egenvärden.
Egenvärde kan allmänt definieras på följande sätt:
Om A är en $n \times n$-matris, så kallas en (icke-noll) vektor $\boldsymbol{x}$ i R^n för en *egenvektor* av A om $A\boldsymbol{x}$ är en skalär multipel av $\boldsymbol{x}$; dvs. $A\boldsymbol{x} = \lambda\boldsymbol{x}$ för någon skalär λ. Skalären λ kallas för *egenvärdet* av A, och $\boldsymbol{x}$ sägs vara egenvektorn som svarar mot λ. [2]

Plotten kan användas för att bestämma antalet komponenter att behålla och visar alltid egenvärdena i en sluttande kurva, från störst till minst. För att besluta antal komponenter att behålla kan man använda sig av Kaiser's regel som helt enkelt en tumregel om att behålla alla komponenter med egenvärden större än 1. [11]

2.4.2 Binära Regressions-träd

Idén är att designa en algoritm som partitionerar träningsmängden till disjunkta (homogena) delmängder. I varje delmängd väljer vi en frekvensparameter λ som beskriver den förväntade frekvensen på den delen av data. Slutligen skattar vi den (okända) förväntade frekvensen på det totala träningsmängden.
För partitioneringen använder vi en binär-träd-algoritm, som börjar med att partitionera träningsmängden i två olika disjunkta delmängder. Dessa delmängder kan sedan var och en partitioneras i två disjunkta delmängder, och så vidare. En bra partition av träningsmängden för vårt prediktionsproblem kommer att beskriva bra val av binära splittar (delningar) av noder och en stoppregel för när inga fortsatta splittar ska göras på en specifik nod och det binära trädet (respektive).

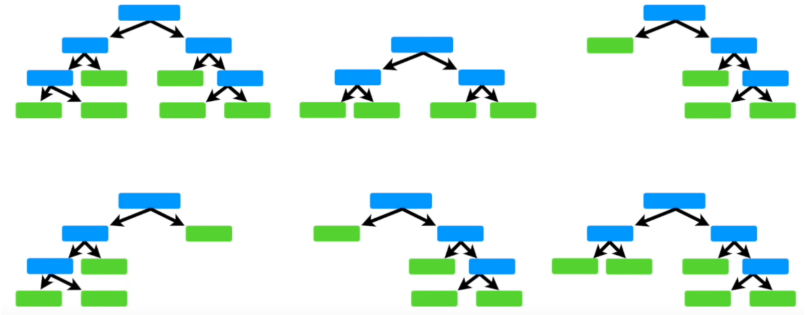


Figur 2. *Illustration av ett möjligt träd där blå noder har splittats och de gröna noderna är löv [19].*

2.4.3 Random forest

Random Forest är en träd-baserad machine learning-algoritm (inom gruppen "ensemble-metoder") och byggstenarna för denna metod är beslutsträd (eller

binära regressions-träd). Inom försäkringsprissättning partitionerar ett beslutsträd en portfölj av försäkringstagare i grupper med homogena riskprofiler baserat på någon egenskap. För varje delgrupp förs en konstant prediktion fram och inkluderar automatiskt solidariteten bland försäkringstagarna i en delgrupp (så länge storleken på denna grupp är tillräckligt stor). Men prediktionsförmågan för sådana enkla träd tenderar att vara ganska låg. Därför betraktar vi en mer komplex algoritm som kombinerar multipla träd i en ensemble, som till exempel Random Forest.



Figur 3. *Illustration av flera olika träd som tillsammans kan utgöra en random forest. På bilden fick vi bara plats med 6 st träd, men det kan alltså vara väldigt många fler [19].*

Bagging

Bootstrap aggregation (Bagging) är en procedur för att reducera variansen av en statistisk inlärningsmetod. Kom ihåg att givet en mängd av n oberoende observationer $Z_1, ..., Z_n$, var och en med variansen σ^2, så ges variansen av väntevärdet $\bar{Z}$ av observationerna av σ^2/n. Med andra ord, att ta genomsnittet av en mängd observationer reducerar variansen. Ett naturligt sätt att reducera variansen för att öka prediktions-noggrannheten av en statistisk inlärnings-metod är att ta många tränings-mängder från populationen, bygga en separat prediktionsmodell genom att använda varje tränings-mängd, och ta genomsnittet av de resulterande prediktionerna. Med andra ord, vi kan beräkna $\hat{f}^1(x), \hat{f}^2(x), ..., \hat{f}^B(x)$ genom att använda B separata tränings-mängder, och ta genomsnittet av dessa för att få en enda statistisk inlärningsmodell med låg varians,

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

Men detta är förstås inte praktiskt eftersom vi i allmänhet inte har tillgång till multipla tränings-mängder. Istället kan vi använda oss av boostrap, genom att ta upprepade stickprov från den (enda) tränings-mängden.
Idén bakom *återsampling* (bootstrap) är att skaffa sig information om t.ex. en parameterskattning genom att dra nya stickprov, inte från den ursprungliga fördelningen, utan från den empiriska fördelningen, som ges av de ursprungliga observationerna, dvs. stickproven dras på nytt från de gjorda observationerna.

För en allmän definition av bootstrap, låt $x_1, ..., x_n$ vara ett stickprov från fördelningen $F(x;\theta)$ och $\theta^*(\boldsymbol{x})$ en skattning av θ.
Metoden kan beskrivas i följande steg:
(1) Bilda ett stort antal, M, nya stickprov av samma storlek, n, som det ursprungliga, men tagna från den empiriska fördelningen F_n istället för F,
(2) Beräkna för vart och ett av stickproven en skattning $\theta_i^*, \quad i = 1, ..., M$,
(3) Beräkna standardavvikelsen, s_{θ^*}, för värdena $\theta_1^*, ..., \theta_M^*$,
(4) Skatta standardavvikelsen $\sqrt{Var(\theta^*)}$ med medelfelet $d(\theta^*) = s_{\theta^*}$.

I denna approach genererar vi B olika bootstrappade tränings-mängder. Vi kan sedan träna vår metod på den b:te bootstrappade träningsmängden för att få $\hat{f}^{*b}(x)$, och sedan slutligen ta genomsnittet av alla prediktioner, för att få

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

Detta kallas *bagging*.
Dessa träd blir djupa och har stor varians, men låg bias. Att ta genomsnittet av dessa B träd reducerar variansen.

Random forests

Random forests, introducerade av Breiman (2001) [3], ger en förbättring över bagging-träd med hjälp av en justering som *dekorrelerar* träden. Precis som med bagging, bygger vi ett antal beslutsträd på bootstrappade stickprov. Men när vi bygger dessa beslutsträd, varje gång en split övervägs i ett träd, så är det ett *slumpmässigt stickprov av m prediktorer* som väljs som split-kandidater från hela mängden av p prediktorer. Splitten tillåts endast att använda en av de utvalda m prediktorerna. Ett nytt stickprov av m prediktorer väljs vid varje split, och typiskt är att vi väljer $m \approx \sqrt{p}$ - dvs. antalet prediktorer som betraktas vid varje split är approximativt ekvivalent med kvadratroten ur det totala antalet prediktorer. Antag att det finns en väldigt stark prediktor i datamängden, tillsammans med ett antal lagom starka prediktorer. Då i samlingen av bagging-träd så kommer alla träd att använda denna starka prediktor i topp-splitten.
Konsekvent, alla bagging-träd kommer se ganska lika ut. Alltså kommer prediktionerna från bagging-träd att vara högt korrelerade. Att ta genomsnittet av många högt korrelerade kvantiteter leder inte till så stor reduktion i variansen som att ta genomsnittet av många okorrelerade kvantiteter. Random forests tar sig runt detta problem genom att tvinga varje split att bara betrakta en delmängd av prediktorer. Därför kommer i genomsnitt $(p - m)p$ av splittarna inte ens att överväga den starka prediktorn. Vi kan tänka på denna process som *dekorrelering*.
Att använda ett litet värde på m när man bygger en random forest kommer att vara till hjälp om vi har ett stort antal korrelerade prediktorer.

För att utföra splitten (recursive binary splitting), väljer vi först prediktorn X_j och delningspunkt s sådana att prediktor-rummet delas upp i områdena $\{X|X_j < s\}$ och $\{X|X_j \geq s\}$ så att vi får den största möjliga reduceringen i RSS (residual sum of squares). (Notationen $\{X|X_j < s\}$ står för det område i

prediktor-rummet i vilket X_j antar ett värde lägre än s.)

Variable-importance

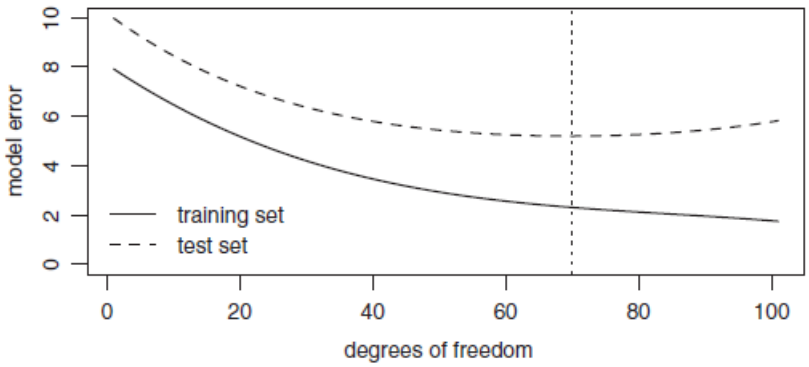
Detta är ett koncept som kan vara svårt att definiera i allmänhet eftersom vikten av en variabel kan bero på dess (möjligen komplexa) interaktion med andra variabler. Random forest-algoritmen skattar vikten av en variabel genom att titta på hur mycket prediktionsfelet ökar när (OOB)-data (se Teori 2.5.3) för den variabeln permuteras medan alla andra variabler lämnas oförändrade. De nödvändiga beräkningarna utförs träd för träd som random forest är konstruerad.
En intressant analys är att studera vilka feature-komponenter (prediktorer) som har valts och hur ofta i Random forest-algoritmen; och med hur mycket dessa splittar har minskat devians-förlusten. Detta ger måttet "Variable importance".

Random forests använder OOB-stickprov för att konstruera ett mått för VI (variable importance), genom att mäta prediktionsförmågan för varje variabel. När det b:te trädet byggs, skickas OOB-stickproven neråt i trädet, och prediktionsnoggrannheten noteras. Sedan permuteras värdena för den j:te variabeln slumpmässigt i OOB-stickproven, och noggrannheten beräknas på nytt. Genomsnittet av avtagandet i noggrannhet (som en följd av denna permutation) över alla träd tas och används som ett mått på VI för variabel j i skogen.

2.5 Modell-validering

2.5.1 GLM, devians-residualer och randomized quantile residuals

För en GLM mäts modell-komplexiteten i termer av frihetsgrader, eller antalet parametrar skattade av modell-anpassnings-proceduren. Varje kontinuerlig variabel vi inkluderar lägger till en frihetsgrad och för en kategorisk variabel läggs en frihetsgrad till för varje nivå som inte är en bascell.
Att öka modell-komplexiteten förbättrar prestationen både på träningsmängden och testmängden - fram till en viss punkt. Efter denna punkt fortsätter prestationen på träningsmängden att förbättras - men på testmängden blir det sämre. Anledningen till detta är att med tillräcklig flexibilitet så är modellen 'fri' att förklara slumpmässigheten i träningsmängdens utfall (brus) tillsammans med den del av data som drivs av systematiska effekter (signal). Bruset i träningsdata kan inte generaliseras till ny data, så i den utsträckning denna information finns i våra modell-skattningar blir det en belastning.



Figur 4. *Illustration av effekten av modell-komplexitet (mäts enligt antal frihetsgrader längs med x-axeln) på modellens prediktionsförmåga (mäts enligt MSE längs med y-axeln) för både träningsmängden och testmängden [5].*

Vårt mål när vi modellerar är att hitta den rätta balansen där vi fångar upp så mycket av signalen som möjligt med minimalt med brus.
Eftersom uppdelningen av data förblir intakt genom hela modellerings-processen är det viktigt att formulera en lämplig datauppdelnings-strategi innan modellbygget påbörjas. Typiska proportioner för denna uppdelning är att 70% av data blir träningsmängd och 30% testmängd eller 60% träningsmängd/40% testmängd (s.38 [5]). Valet av procentandelar innebär en kompromiss. Mer data tillgänglig för träningsmängden kommer att göra det lättare att upptäcka mönster i data. Men om testmängden blir för liten blir den slutliga bedömningen av modellen osäker.

Ett användbart och viktigt medel för att se hur väl en specificerad modell passar till data är genom visuell inspektion av *residualerna*, eller mått på avvikelserna för individuella datapunkter från deras predikterade värden. Devians-residualer är ofta lämpligare inom GLM än rå-residualer och de representerar de individuella observationernas bidrag till deviansen D. Mer specifikt, definieras de som kvadratrötterna (med tecken) för enhetsdevianserna. Alltså devians-residualerna är analoga med traditionella residualer när de är kvadrerade. Till exempel, för Poissonfördelningen definieras devians-residualerna som:

$$r_i = \text{sign}(y_i - \hat{\mu}_i) \cdot \sqrt{(}2\omega_i[y_i \cdot log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]). \tag{2.12}$$

I en väl-anpassad modell kan vi förvänta oss att devians-residualerna har följande egenskaper:
- De följer inget förutsägbart mönster. Residualer ska vara den slumpmässiga (eller icke-predikterbara) delen av data. Om vi upptäcker något sätt som residualerna kan predikteras på så kan vi fortfarande förbättra vår modell.
- De är normalfördelade, med konstant varians. Eftersom devians-residualer har justerats för den underliggande fördelningen, kan man förvänta sig detta. Signifikanta avvikelser från normalfördelningen eller homoskedasticitet kan indikera att den valda fördelningen är felaktig.
För diskreta fördelningar (såsom Poisson och Negativ binomial) kommer deviansresidualerna troligen inte att följa en normalfördelning. Detta beror på att deviansresidualerna inte justerar för diskretheten; det stora antalet observationer

som har samma responsvärde orsakar att residualerna klustrar.
En möjlig lösning är att använda så kallade 'randomized quantile residuals' (s. 66 [5]), som har likande egenskaper som deviansresidualerna, men lägger till slumpmässighet till de diskreta punkterna.
Låt $F(y; \mu, \phi)$ vara fördelningsfunktionen av responsvariabelns fördelning. Om F är diskret låt $a_i = \lim_{y \uparrow y_i} F(y; \hat{\mu}_i, \hat{\phi})$ och $b_i = F(y_i; \hat{\mu}_i, \hat{\phi})$. Vi definierar randomized quantile residuals för y_i med $r_{q,i} = \Phi^{-1}(u_i)$, där u_i är en likformig slumpvariabel på intervallet $(a_i, bi_i]$ [4].

2.5.2 Logistisk regression och ROC-kurvor

En logistisk modell predikterar som tidigare nämnt sannolikheten för en händelses inträffande, men den sannolikheten kommer att behöva översättas till en binär prediktion av förekomst/ej förekomst för att kunna ta ett beslut. Till exempel, antag att vi bygger en modell för att upptäcka bedrägerier; för varje ny skada ger modellen en sannolikhet att det är bedrägeri. Baserat på denna prediktion måste vi ta beslut om hur man vidare ska undersöka skadan. Vi kan göra ett sådant beslut genom att välja en specifik sannolikhets-nivå, säg 50%, ovan vilken vi undersöker skadorna och under vilken vi inte gör det.
Då är följande fyra utfall möjliga:
1.) Modellen predikterar att skadan innehöll bedrägeri och det stämmer (true positive, TP).
2.) Modellen predikterar att skadan innehöll bedrägeri och det stämmer inte (false positive, FP).
3.) Modellen predikterar att skadan inte innehöll bedrägeri och det stämmer (true negative, TN).
4.) Modellen predikterar att skadan inte innehöll bedrägeri och det stämmer inte (false negative, FN).

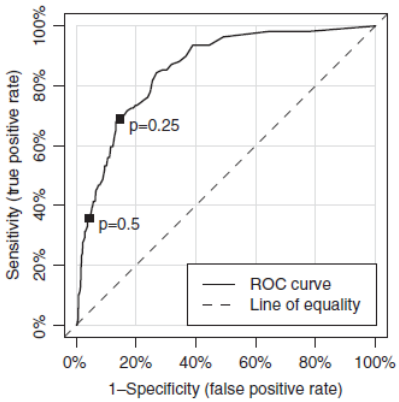
Vi kan sedan gruppera observationerna enligt de fyra kombinationerna ovan av faktiska och predikterade utfall, och räkna antalet observationer som hamnar i varje grupp. Vi kan visa dessa resultat i en så kallad *confusion matrix*:

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figur 5. *Illustration av en confusion matrix [5].*

Kvoten av TP:s och totalt antal positiva händelser kallas för *sensitivity*. Kvoten av TN:s och totalt antal negativa händelser kallas *specifity*.
Man skulle kunna fånga upp fler bedrägerier genom att sänka tröskeln till 25%, men då till priset av att en ökad frekvens av FP:s.

Ett lämpligt grafiskt verktyg för att evaluera modellen är en ROC- (*receiver operating characteristic*) kurva, vilken konstrueras genom att plotta FP-frekvenserna längs med x-axeln och TP-frekvenserna längs med y-axeln för olika tröskelvärden i intervallet $[0, 1]$.



Figur 6. *ROC-kurva för en exempel bedrägeri-modell [5].*

En modell som ger prediktioner som inte är mycket bättre än slumpen kommer att ge TP:s och FP:s i samma proportioner som den totala blandningen av positiva och negativa i data, oavsett vilket tröskelvärde som väljs. För en sådan modell kommer ROC-kurvan att följa likhetslinjen ('line of equality' i figuren ovan).
En modell med hög prediktionsförmåga kommer att ge TP:s med högre frekvens än FP:s, vilket resulterar i en ROC-kurva som är högre än likhetslinjen. Förbättrad noggrannhet kommer att öka kurvans avstånd från likhetslinjen.

2.5.3 Random Forest och Out-of-Bag

Det visar sig att det finns ett väldigt straight-forward sätt att skatta testfelet hos en bagging-modell. Nyckeln till bagging är att träd upprepade gånger anpassas till bootstrappade delmängder av observationerna. Man kan visa att i genomsnitt, så använder varje bagging-träd omkring 2/3 av observationerna. Den kvarvarande tredjedelen av observationerna som inte används för att anpassa ett givet bagging-träd refereras till som *out-of-bag* (OOB)-observationer. Vi kan prediktera responsen för den i:te observationen genom att använda alla träd som hade just den observationen i OOB. Detta kommer att ge ungefär B/3 prediktioner för den i:te observationen. För att få ut en enda prediktion för den i:te observationen, kan vi ta genomsnittet av dessa predikterade responser (regression). Detta leder till en enda OOB-prediktion för den i:te observationen. En OOB-prediktion kan fås på detta sätt för var och en av de n observationerna, från vilka den totala OOB-MSE (Mean squared error) kan beräknas:

$$\mathrm{MSE}_{OOB} = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{y}_i^{OOB}\}^2, \tag{2.13}$$

där $\hat{y}_i^{OOB}$ är genomsnittet av OOB-prediktionerna för den i:te observationen. Det resulterande OOB-felet är en giltig skattning för testfelet för den bagging-modellen, eftersom responsen för varje observation predikteras genom att bara använda de träd som inte anpassades med just den observationen.

2.5.4 Validation set approach och Actual vs. Predicted

Testfelet är det genomsnittliga fel som resulterar från att använda en statistisk inlärningsmetod för att prediktera responsen för en ny observation - dvs. ett mått som inte används när man tränar metoden. Givet ett dataset, så är en specifik metod rättfärdigad om den resulterar i ett lågt testfel (test error rate). Testfelet kan enkelt beräknas om en testmängd finns tillgänglig. Oturligt nog, så är detta inte alltid fallet. I kontrast, så kan träningsfelet (training error rate) lätt räknas ut genom att tillämpa den statistiska inlärningsmetoden till observationerna som användes till dess träning. Men skillnaden mellan testfelet och träningsfelet är ofta ganska stor, och speciellt så kan den den senare dramatiskt underskatta den förra.

Validation set approach är en mycket enkel strategi för att göra just detta. Den innebär att man slumpmässigt delar in den tillgängliga observationsmängden i två delar, en träningsmängd och en testmängd. Modellen anpassas på träningsmängden, och den anpassade modellen används för att prediktera responserna i testmängden.
Den här metoden är väldigt enkel och lätt att implementera. Men den har två potentiella nackdelar:
(1) testfel-skattningen kan variera mycket, beroende på precis vilka observationer som inkluderats i träningsmängden och testmängden.
(2) Endast en delmängd av observationerna används för att anpassa modellen. Eftersom statistiska metoder tenderar att prestera sämre när de tränas på färre observationer, så antyder detta att träningsfelet tenderar att överskatta testfelet för modellanpassningen på hela datasetet.

Ett enkelt sätt att jämföra olika modellers prediktionsförmåga är att göra en så kallad Actual vs. predicted-plot (scoring) på test-data. Om modellen i fråga har god prediktionsförmåga borde vi få ett linjärt samband.
Det är ofta nödvändigt att aggregera data före plotten i enlighet med storleken på data. En vanlig approach är att gruppera data i många mindre grupper och beräkna väntevärdet för varje grupp av predikterade och faktiska responsvariabelvärden och sedan plotta dessa mot varandra.

3. Data

3.1 Försäkringsdata

Data som vi baserar analysen på är historiska data för ICA försäkrings motoraffär mellan 2015-2019.
Vi har från början totalt 611 052 observationer, där varje observation är en försäkringsperiod, vanligen ett år. Det förekommer dock flera fall där exponeringen är bara ett par dagar och även längre än ett år. I det första fallet så beror det oftast på att kunden har gjort något tillägg eller liknande i sin försäkring och då blir till en ny observation. De durationer som överstiger ett år litegrann är oftast sådana att kunden har struntat i att betala och då gäller försäkringen ett antal veckor till innan den slutligen sägs upp och eftersom det är ytterligare riskexponeringstid har detta räknats med i data.

Det förekommer en del överlapp mellan de tre olika grenarna trafik, delkasko och vagn. Men vi gör en analys per gren och har för: Vagnskada: 111386 observationer, Delkasko: 237144 observationer och Trafik: 262522 observationer.
Vi behåller nollskadorna i data eftersom dessa likväl är skador även om de inte genererar några utbetalningar. De observationer som innehåller nollskador utgör dock bara någon procent av datamaterialet. Vi intresserar oss här bara för kvantiteten för att bedöma huruvida någon är risk-aversiv eller inte.
Vi bortser från variabeln med skadekostnader eftersom vanligt är att anta att skadefrekvensen och medelskadan är oberoende. Vi har även tagit bort observationer som innehåller rena orimligheter.
Från början har jag tillgång till information om fyra olika bank-engagemang, men för att få ett helhetssvar har jag skapat en ny binär variabel, Bank, som bara svarar på om en viss kund har ett eller flera bankengagemang ('Ja') eller inget bankengagemang ('Nej').

	min dur	max dur	mean dur	andel skador	andel bank-kunder
Vagnskada	2,00137	528,362	203,9672	0,0735	0,595003
Delkasko	2,00137	557,382	184,4656	0,063	0,5511461
Trafik	2,00137	557,382	183,7169	0,0344	0,5227143

Figur 7. *Egenskaper hos datamängden för de tre grenarna. Med min-,max-värde och väntevärde för durationen (i dagar), andel skador och andel bank-kunder.*

3.2 Uppdelning av data

Innan man påbörjar modelleringen är det viktigt att dela in tillgängliga data i två grupper. En av dessa grupper kallas för *träningsmängden*. Den kommer vi att använda för alla modell-uppbyggnads-steg. Den andra gruppen kallas för *testmängden* och kommer att användas för att bedöma modellens prediktionsförmåga/prestation.
En anledning till att vi gör detta är att om vi försöker testa prediktionsförmågan för någon modell på samma data som användes för att bygga upp modellen, så kommer vi att få överoptimistiska resultat.
Vi delar upp data i en träningsmängd (70% av data) och en testmängd (30% av data), båda slumpmässigt valda från den ursprungliga datamängden.

3.3 Oberoende observationer

Som beskrivet i teori-delen (2.3) så måste observationerna som modellen bygger på vara oberoende. Det är för att vi endast vill modellera hur de förklarande variablerna i modellen påverkar utfallet av responsvariabeln och inte att denna effekt blir svår att urskilja pga. hur en annan observation påverkar utfallet. Det skulle till exempel vara problematiskt om båda individer i en och samma bilolycka anmäler varsin skada och har samma försäkringsbolag.

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.

3.4 Variabler

Vi har totalt 13 prediktorer som kan tänkas ha betydelse för vår frågeställning (se Figur 8).

	Variabeltyp
Variabel A	heltalsvärd
Variabel B	kategorisk
Variabel C	heltalsvärd
Variabel D	kategorisk
Variabel E	kategorisk
Variabel F	heltalsvärd
Variabel G	kategorisk
Variabel H	heltalsvärd
Variabel I	kategorisk
Variabel J	kategorisk
Variabel K	kategorisk (intervall)
Variabel L	kategorisk
Bank	binär
Antal skador	kategorisk
Duration	kontinuerlig
Skada (log reg)	binär

Figur 8. *Tabell med pseudonymiserade variabler och deras variabeltyper. Även de icke-pseudonymiserade variablerna längst ner i tabellen, där 'Skada' är en binär variabel som vi skapar för den logistiska regressions-modell.*

3.5 Dataanalys

Vi vill utforska data genom att undersöka:
(i) samband mellan respons och potentiella prediktorer; och
(ii) samband mellan potentiella prediktorer.
Resultatet av (i) föreslår variabler eller riskfaktorer för modellen, och deras troliga effekt på responsen. (ii) belyser vilka prediktorer som är associerade. Starkt associerade prediktorer inkluderas i modellen med försiktighet.

Vid jämförelse av två variabler finns det olika metoder beroende på variabeltyp. Sambandet mellan två kontinuerliga variabler utforskas med hjälp av en scatterplot. En frekvenstabell är det vanligaste sättet att visualisera på när man undersöker samband mellan två kategoriska variabler. Boxplottar är lämpliga för att undersöka en kontinuerlig variabel mot en kategorisk. Ofta blir det svårt att uttyda något från rådata eftersom det ofta kan finnas en extrem skevhet i dessa. Bättre är att använda logaritmerade data (för den kontinuerliga variabeln).

Inga specifika samband upptäcktes i någon av datamängderna med hjälp av metoderna ovan. Då vi har ganska många variabler väljer vi att inte presentera alla dessa plottar i denna rapport.

Vi tittar även efter multikolinjäritet med hjälp av VIF. Enligt gäller det att GVIF > 10 tyder på multikolinjäritet. Resultatet enligt figur 9, tyder alltså på att ingen av våra variabler innehåller multikolinjäritet. Detta gäller alla grenarna.

	GVIF	Df	GVIF^(1/2*Df)
A	1,425912	3	1,06E+00
B	1,289855	5	1,03E+00
Bank	1,058919	1	1,03E+00
C	1,534842	20	1,01E+00
D	2,010365	19	1,02E+00
E	1,043296	2	1,01E+00
F	1,585674	2	1,12E+00
G	1,238113	4	1,027058
H	1,278584	2	1,06E+00
I	1,103331	4	1,012368
J	1,062406	3	1,01E+00
K	1,062615	2	1,02E+00
L	1,201319	2	1,046923

Figur 9. *Tabell över de GVIF-värden vi får fram för alla variabler.*

4. Metod och resultat

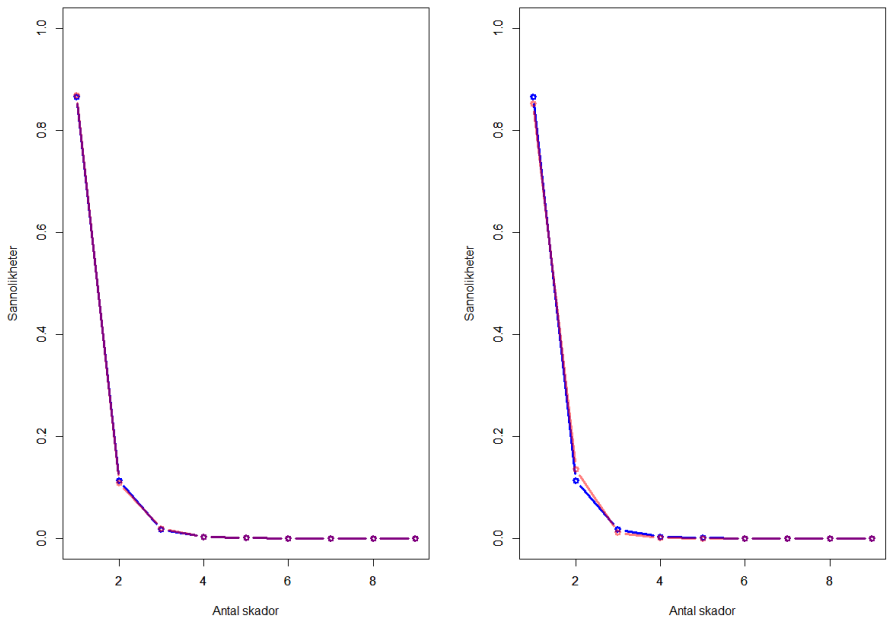
Vi ska prova tre olika metoder: GLM (multiplikativ modell), logistisk regression och Random Forest. För de båda metoderna GLM och logistisk regression använder vi från R-package MASS funktionerna 'glm.nb','glm' och för Random forest använder vi R-package 'randomForest'.
Här kommer vi, med anledning av att Vagnskade-grenens datamängd är den med lägst antal observationer, att använda datamängden för Vagnskade-grenen för att beskriva metoden. Resultat presenteras dock för alla tre grenarna.

4.1 Multiplikativ GLM-modell

När länkfunktionen för en GLM specificeras att vara den naturliga logaritmfunktionen (dvs. $g(x) = ln(x)$), så produceras en multiplikativ värderings-struktur, ty

$$\begin{aligned} \ln(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \quad \Leftrightarrow \\ \mu_i = \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) = e^{\beta_0} \cdot ... \cdot e^{\beta_p x_{ip}} \end{aligned} \tag{4.1}$$

Eftersom skadefrekvensen är så starkt korrelerad med risk-aversiviteten, så utgör den kontinuerliga variabeln skadefrekvens vår responsvariabel i denna modell.
Vi ska jämföra den empiriska fördelningen mot Poisson- (Po(λ)) och Negativa Binomial-fördelningen (NegBin(α, β)) för att se om någon av dom verkar passa våra data. Eftersom väntevärdet i Poissonfördelningen är λ, ger detta direkt skattningen $\hat{\lambda} = \bar{x}$. Om vi aggregerar data för att få en observation per kund kan vi låta n_k beteckna antalet observationer som har antal skador lika med k samt betecknar antalet observationer med n. Det vi jämför är alltså sannolikhetsfördelningarna för den empiriska fördelningen, dvs. $p_k = n_k/n$, för Poissonfördelningen, dvs. $p_k = \frac{\lambda^k}{k!}e^{-\lambda}$, och för Negativa Binomialfördelningen $p_k = \frac{\Gamma(\alpha+k}{\gamma(\alpha)k!}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{k}$. β- och α-parametrarna skattar vi med hjälp av det empiriska väntevärdet och den empiriska variansen enligt s.49 [9].



Figur 10. *Resultat för jämförelser av empiriska fördelningen (blått) och negBin/Poisson (rött). I grafen till vänster representerar den röda linjen negativa binomialfördelningen och i grafen till höger representerar den röda linjen Poisson-fördelningen.*

Tittar vi noga i figuren så ser vi i den vänstra grafen att Negativ Binomial-fördelningen passar allra bäst, även om Poisson skulle ha kunnat användas med gott resultat också. Vi fortsätter med negativa binomial-fördelningen.

Vi använder funktionen glm.nb() i R för att göra en multiplikativ glm-analys där responsen (skadefrekvensen) antas vara negativt binomialfördelad.
Eftersom durationen är en kontinuerlig variabel ska den logaritmeras (se Teori 2.3.2). Bascellerna väljs efter ”störst population” (flest antal observationer).

Till att börja med vill vi göra hierarkiska tester för de olika variablerna för att se om någon av dom kan uteslutas. Nollhypoteserna är då att premieargumentet i fråga inte ger tillräckligt med effekt på skadefrekvensen och därför kan tas bort från modellen.
I första iterationen blir resultatet enligt Figur 11 (där jag har pseudonymiserat alla prediktorer utom Bank med bokstäver).

	df	LR-statistika	$\chi^2(df)$ p-värde
A	3	22,95712	4,12E-05
B	5	262,7992	9,84E-55
Bank	1	43,22842	4,87E-11
C	20	100,6413	9,67E-13
D	19	96,38791	2,40E-12
E	2	50,62828	1,01E-11
F	2	32,14666	1,05E-07
G	4	14,52326	0,005799349
H	2	74,92461	5,37E-17
I	4	8,077437	0,0887824
J	3	93,54309	3,80E-20
K	2	33,04771	6,66E-08
L	2	4,054082	0,1317247

Figur 11. *Tabell över resultatet från den första hierarkiska testomgången.*

Vi gör detta hierarkiska test för varje premieargument var för sig och med 0.05 signifikansnivå, och det premieargument med sämst signifikans väljer vi att först studera klass-konfidensintervallen för, vilket i detta fall är premieargument L. Som synes i figur A1 i Appendix så är alla klasser av premieargument L (utöver bascellen) icke-signifikanta, dvs. konfidensintervallen innehåller 1, och kan därmed lika gärna höra till bascellen. Vi beslutar oss därför för att utesluta premieargument L.
Efter ett nytt hierarkiskt test får vi resultat enligt figur 12.

	df	LR-statistika	$\chi^2(df)$ p-värde
A	3	10,54804	1,44E-02
B	5	160,7606	6,81E-13
Bank	1	19,15492	1,21E-05
C	20	197,0213	4,37E-31
D	19	97,1597	1,74E-12
E	2	47,13376	5,82E-11
F	2	27,09381	1,31E-06
G	4	5,99956	0,1991811
H	2	31,74459	1,28E-07
I	4	4,810314	0,30732
J	3	94,42222	2,46E-20
K	2	30,38092	2,53E-07

Figur 12. *Tabell över resultatet från den andra hierarkiska testomgången, undantaget variabel L.*

På samma sätt som tidigare väljer vi att titta på konfidensintervallen för premieargument I då den har sämst signifikans. Som synes i figur A2 i Appendix, så är alla klasser av premieargument I (utöver bascellen) icke-signifikanta. Vi utesluter premieargument I.
Efter ett nytt hierarkiskt test får vi resultat enligt figur 13.

	df	LR-statistika	$\chi^2(df)$ p-värde
A	3	10,43682	1,52E-02
B	5	162,1389	3,46E-33
Bank	1	19,93119	8,03E-06
C	20	196,0245	6,87E-31
D	19	97,91133	1,28E-12
E	2	46,98546	6,27E-11
F	2	26,33346	1,91E-06
G	4	5,903002	0,2065102
H	2	32,41658	9,14E-08
J	3	95,75689	1,27E-20
K	2	31,67502	1,32E-07

Figur 13. *Tabell över resultatet från den tredje hierarkiska testomgången, undantaget variabel I och L.*

Eftersom premieargument G då har sämst signifikans och alla klasser av premieargument G (utöver bascellen) icke-signifikanta (Se figur A3 i Appendix), beslutar vi oss även här för att utesluta premieargument G.
Ett nytt hierarkiskt test enligt figur 14 visar att alla kvarvarande premieargument nu är signifikanta.

	df	LR-statistika	$\chi^2(df)$ p-värde
A	3	10,43893	1,52E-02
B	5	163,1814	2,08E-33
Bank	1	19,89596	8,18E-06
C	20	217,1235	4,47E-35
D	19	98,85219	8,64E-13
E	2	45,57662	1,27E-10
F	2	27,88236	8,82E-07
H	2	32,07149	1,09E-07
J	3	95,43907	1,49E-20
K	2	30,60047	2,27E-07

Figur 14. *Tabell över resultatet från den fjärde hierarkiska testomgången, undantaget variabel I,G och L.*

Nu ska vi titta på konfidensintervallen för de kvarvarande premieargumenten och deras klasser. Vi vill se om det finns klasser som kan slås ihop (se Figur A3 i Appendix). Till att börja med kan vi i figuren se att det finns en del tariffceller som innehåller väldigt få observationer (population). Detta i sig gör att vi inte säkert kan lita på de resultat vi får utifrån dessa, på grund av det tidigare nämnda att GLM tilldelar alla tariffceller lika hög trovärdighet (kredibilitet). De klasser som vi här noterar med mycket få observationer är C1,C2,C3,H2 och K2. Vi slår ihop dessa klasser med lämplig annan klass inom samma premieargument så att C4 nu består av de observationer som tidigare var C1,C2,C3 och C4, och H3 består nu av de observationer som tidigare var H2 och H3, och till sist så består K1 nu av de observationer som tidigare utgjorde K1 och K2.
Tittar vi nu på konfidensintervallen kan vi se att premieargument D har så många som 14 (!) icke-signifikanta klasser. Men eftersom premieargumenten har olika många klasser från början börjar vi med det som har störst andel icke-signifikanta klasser (vilket i vårt fall ändå betyder att vi börjar med D). Vi slår ihop de andra klasserna med D20 ett och ett ordnade efter hur nära konfidensgränserna 1 ligger (längst avstånd först, dvs. den klass som är troligast att sammanfalla med bascellen). Vi itererar detta förfarande tills vi har en stabil tabell med enbart signifikanta klasser (se Figur A5 i Appendix).

Vi gör även ett nytt hierarkiskt test och alla premieargument (inkl. Bank-variabeln) nu mycket signifikanta.
Vi har nu alltså kommit fram till att Bank-variabeln inom Vagnskade-grenen har påverkan på skadefrekvensen. Men vi vill också veta åt "vilket håll"(positivt eller negativt) Bank-variabelns klasser 'Ja' och 'Nej' påverkar skadefrekvensen. Eftersom vi har satt BankJa som bascell (dvs. har relationstal 1) och relationstalet för klassen BankNej i figur A5 i Appendix är 1.1118402, betyder det att väntevärdet i BankNej-cellen är ca 11.2% högre än väntevärdet i BankJa-cellen. Det vill säga den förväntade skadefrekvensen är lägre för de kunder som har ett bank-engagemang. Med andra ord blir slutsatsen alltså att inom grenen vagnskada så är kunder med bankengagemang mer risk-aversiva än andra kunder.

För grenarna Delkasko och Trafik blir slutsatsen densamma men i var och en av grenarna så är väntevärdet i BankNej-cellen 4.93% respektive 17.76% högre än väntevärdet för respektive grens BankJa-cell.

4.2 Logistisk regression

I det här fallet är vår responsvariabel istället en binär variabel som mäter sannolikheten för att en skada inträffar resp. inte inträffar. Sannolikheten för att en skada inträffar ska jämföras för bank-kunderna resp. icke bank-kunder.

Eftersom logistisk regression är ett specialfall av en glm-modell är mycket av förberedelserna demsamma som i förra avsnittet, men istället för att använda oss av offset som i den multiplikativa GLM så gör vi en egen ny variabel kallad Skada, som bara innehåller 'Ja' eller 'Nej' och tar alltså inte hänsyn till hur många skador per kontrakt det är. Dessa värden på Skada viktas sedan enligt durations-variabeln (dvs. hur lång tids riskexponering kontraktet har haft). Enligt samma förfarande som för GLM tidigare itererar vi oss fram till en stabil modell där alla nivåer är signifikanta (se Figur A6 i Appendix).

Tittar vi på relationstalen i figur A6 i Appendix, så kan vi också se att relationstalet för nivån Nej"i Bank-variabeln är 0.9068678. Det betyder att väntevärdet i BankNej-cellen är ca 9.3% lägre än väntevärdet i BankJa-cellen. Det vill säga den förväntade sannolikheten för skada är lägre för de kunder som inte har ett bank-engagemang. Med andra ord blir slutsatsen alltså att inom grenen vagnskada så är kunder utan bankengagemang mer risk-aversiva än andra kunder.

För grenarna Delkasko och Trafik blir slutsatsen densamma men i var och en av grenarna så är väntevärdet i BankNej-cellen 0.93% respektive 11.08% lägre än väntevärdet för respektive grens BankJa-cell.

4.3 Random Forest

4.3.1 Implementation i R

randomForest()-funktionen i R implementerar Breiman's random forest algoritm [18]. Variable importance beräknas som den genomsnittliga ökningen av kvadraten av out-of-bag residualer då variabeln är permuterad. För vår frågeställning är vi mest intresserade av detta mått.

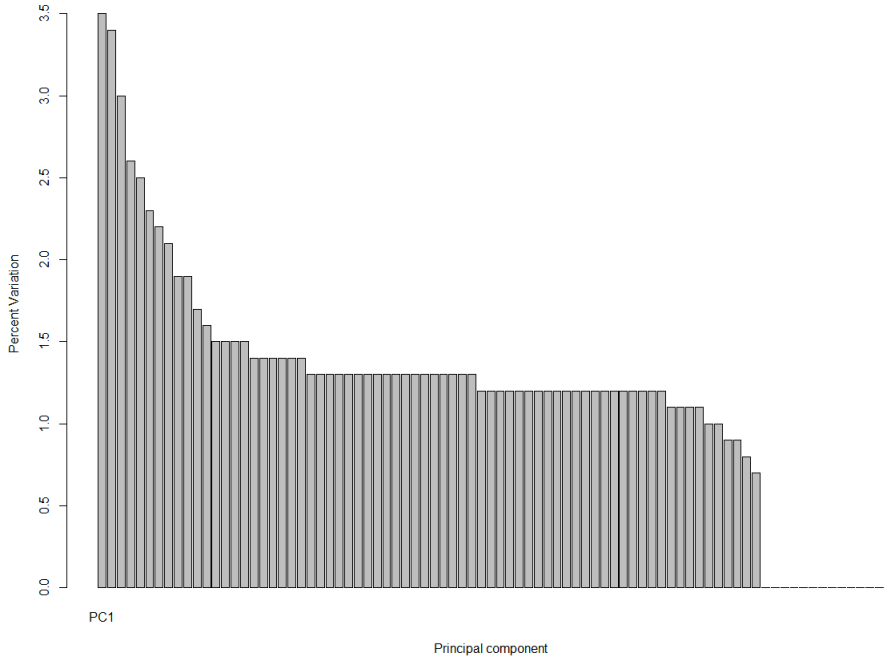
4.3.2 Metod

Eftersom risk-aversivitet hos kunderna kan bero på många saker finns det inget enkelt sätt att klassificera data. Att bara klassificera enligt Skada: Ja/Nej som i Logistisk regression gör att vi tappar informationen om antalet skador i

förhållande till durationen. I Logistisk regression får vi ut den förväntade sannolikheten för en skada, men i detta fall skulle vi alltså bara få ut vilken delmängd av populationen som har flest skador. Vi kommer alltså att göra en Random Forest för regression.

Vi skapar en ny variabel för skadefrekvens och eliminerar duration- och skadeantalsvariablerna. De kategoriseringar som vi skapade för vissa variabler i GLM är nu återställda.

Vi vill först göra en kontroll av om det går att förenkla modellen, så kallad dimensions-reduktion. En sådan metod är tidigare omnämnda PCA (Principal component analysis). För att kunna utföra en sådan analys behöver alla variabler vara numeriska. Men då vi har flera kategoriska variabler gör vi här om varje kategori till en dummy-variabel med värdena 0 och 1 som då blir nya variabler (Så kallad 'One Hot encoding'). Det innebär att vi nu har 82 prediktorer istället för 13.



Figur 15. *Scree plot. Visar hur stor procentandel av den totala variansen som beskrivs av principal komponent nr x (x-axeln).*

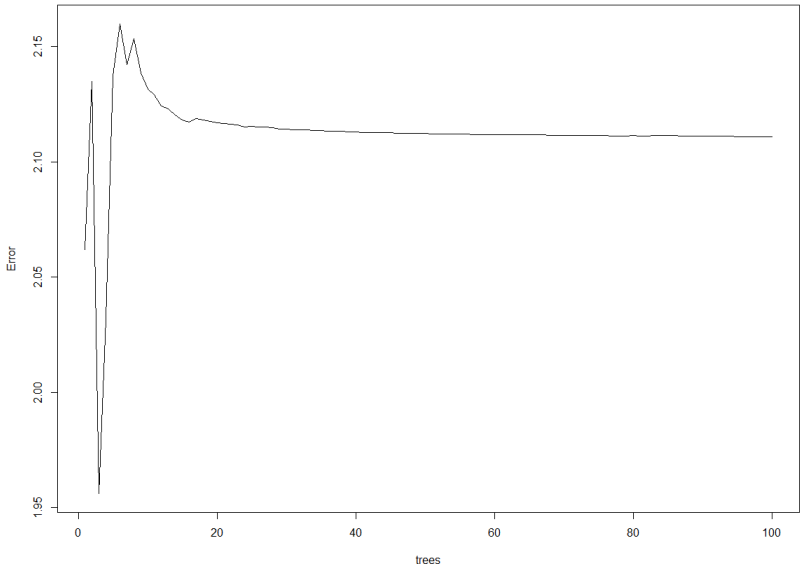
I figuren ser det dock ut som om vi inte har särskilt mycket korrelation mellan variablerna och att de flesta av dom inte kan uteslutas hur som helst. Även om vi skulle använda oss av Kaiser's regel om att bara behålla principal-komponenter med egenvärde > 1, så blir det ändå hela 51 principal-komponenter kvar. PCA är mest användbart då det är några få principal komponenter som blir över.
Enligt [3] så kan variablerna vara både kontinuerliga och/eller kategoriska med

random forest, men för PCA var det nödvändigt att använda 'One Hot Encoding' som beskrivet ovan. Men eftersom PCA inte verkade ge upphov till någon tydlig förbättring återgår vi till hur data såg ut innan med 13 prediktorer igen.

Eventuellt kan man använda subset selection istället i ett försök att reducera antalet prediktorer, men problemet med denna metod är att den tittar på alla 2^p möjliga modeller av p variabler. Så i detta fall blir det $2^{13} = 8192$ modeller som ska gås igenom, dvs. det blir beräkningstungt. Man kan också använda Ridge-, Lasso- eller Elastic-net-regularization, men det blev ingen nämnvärd skillnad.

Vi fortsätter med att använda en Random Forest-algoritm på data. Vid koll av vilket antal (m) variabler algoritmen ska ha att välja mellan per split som ger lägst fel, så kan vi se att i vårt fall blir felet (OOB MSE ekvation (2.13)) lägre ju mindre värde på m vi har, vilket inte stämmer med tumregeln att $m = \sqrt(p)$, om p är antalet prediktorer vi har. Alltså vi får det lägsta felet när det bara finns en variabel att välja på i varje split. Eftersom vi har 13 variabler borde det enligt tumregeln alltså vara $\sqrt(13) \approx 3.6 \approx 4$ variabler per split. Men att bara använda en slumpmässig variabel i varje split kan bli alltför slumpmässigt. Vi fortsätter med att ha 4 variabler för algoritmen att välja på i varje split ändå.

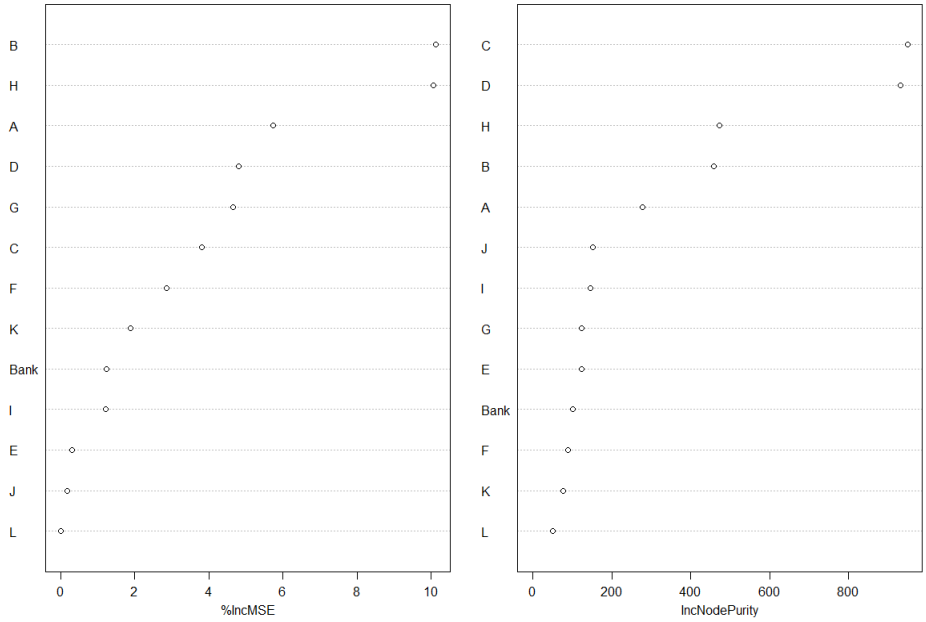
Tittar vi på en plot av felet och antalet träd (figur 16), så kan vi se att vi har konvergens redan efter ca 20 träd.



Figur 16. *Visar hur OOB-felet (y-axeln) minskar med antalet träd (x-axeln), med stoppregel nodstorlek=1000. Gren: Vagnskada.*

För den implementation i R som vi använder beräknas detta med hjälp av justerad R^2 som är nästan samma sak som R^2 (se ekvation (2.6)) med den skillnaden att den förra tar hänsyn till vilka variabler som bidrar till prediktionen genom att dividera med antalet frihetsgrader.

En enkel tolkning av det mycket låga procenttal vi erhåller, är att vi lika gärna kunde prediktera varje stickprov till att vara ekvivalent med det totala väntevärdet. Det vill säga modellen fungerar inte särskilt bra.
Eftersom skadefrekvensen i våra data har en ganska lång svans prövar vi också att ta bort de observationer som bildar denna (outliers) eftersom random forest inte alltid hanterar detta bra, men vi fick liknande resultat ändå, så ingen större förbättring.



Figur 17. *Variable importance. Till vänster: variable importance baserad på den förväntade minskningen av prediktionsnoggrannhet av OOB-stickproven när en variabel exkluderas från modellen. Till höger: visar också variable importance men istället baserad på den totala reduceringen i node impurity som resulterar från splittar över variabeln i fråga, genomsnittligt över alla träd.*

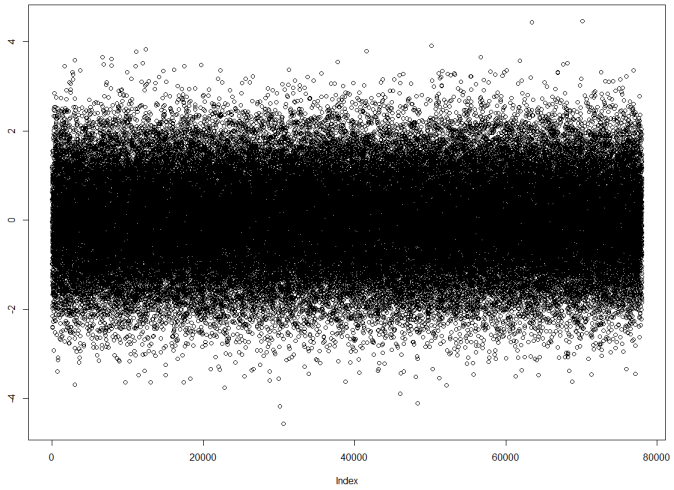
I fallet med regressionsträd mäts node impurity med träningsmängdens RSS (residual sum of squares).

5. Validering och jämförelse

För att validera de tre modellerna använder vi olika mått som lämpar sig väl för varje metod för sig. Sedan ska vi jämföra testfelen och prediktionsförmågan hos alla modellerna. För varje gren använder jag samma samplade tränings- och test-mängder.

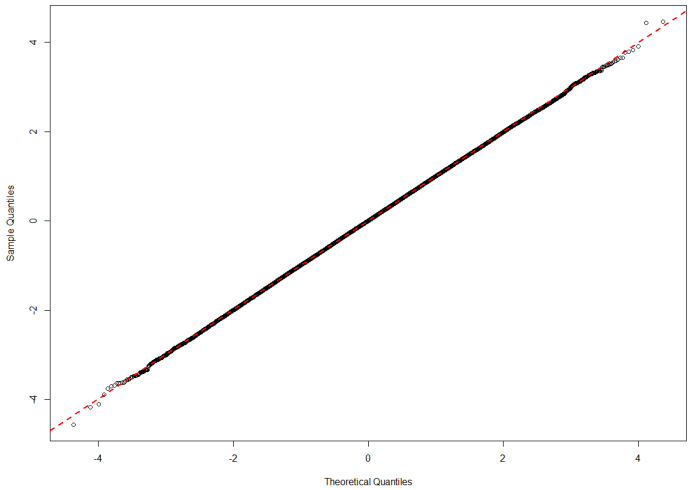
5.1 Validering av Multiplikativ GLM

Vi använder oss av randomized quantile residuals (RQ-residualer) eftersom vi har antagit den diskreta fördelningen negativ binomial för vår respons. Till att börja med tittar vi efter eventuella mönster hos dessa residualer med hälp av en scatterplot (se figur 19).



Figur 18. *En scatterplot av RQ-residualerna.*

Alla residualer håller sig i ett horisontellt band kring noll och uppvisar därmed inga predikterbara tendenser. Som synes i plotten nedan kan RQ-residualerna helt klart sägas komma från en normalfördelning.



Figur 19. *En QQ-plot av RQ-residualerna och normalfördelningens kvantiler.*

Ett liknande resultat erhåller vi för de andra två grenarna.

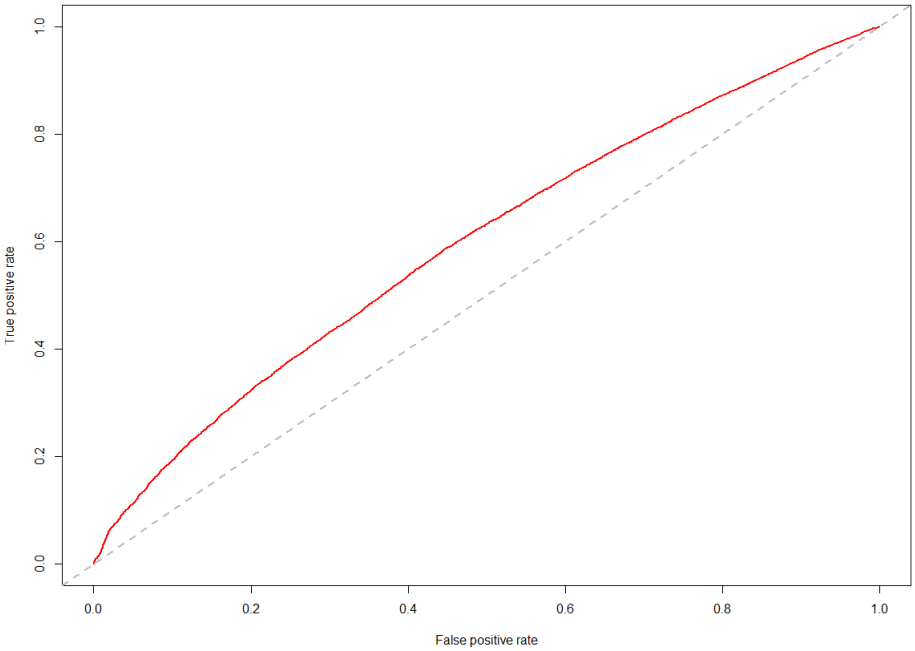
5.2 Validering av Logistisk regression

Vi predikterar sannolikheten att det inträffar en skada, givet värdet på prediktorerna. Sannolikheterna blir på formen $P(Y = 1|X)$. För att kunna göra en prediktion av om det kommer att inträffa en skada under en specifik försäkringsperiod, måste vi konvertera dessa predikterade sannolikheter till klasser, Ja och Nej. Vi skapar en vektor med klass-prediktioner baserad på huruvida den predikterade sannolikheten för en skada är större än eller mindre än 0.5.

```
> table(glm.pred,train$Skada)

glm.pred    Ja   Nej
     Nej  5714 72256
> mean(glm.pred==train$Skada)
[1] 0.9267154
```

Figur 20. *Confusion matrix för vagnskade-grenen.*

Som vi kan se så får vi bara en rad i matrisen eftersom det är så stor sannolikhet att det inte sker någon skada att alla prediktioner klassificerades som Nej".
Vi konstruerar en ROC-kurva:



Figur 21. *ROC-kurva. Vagnskada.*

Att ROC-kurvan inte är så långt ifrån likhetslinjen tyder på att modellens prediktionsförmåga är lite bättre än slumpen, men inte mycket.

5.3 Validering av Random Forest

Ur sammanfattningen nedan för Vagnskade-grenen som vi får från Random Forest i R kan vi se att variansen är förklarad till 0.45% och MSR (mean squared residuals) är 2.628379.

```
Call:
 randomForest(formula = skfreq ~ ., data = Hull, mtry = 4,
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 4

          Mean of squared residuals: 2.628379
                    % Var explained: 0.45
```

Figur 22. *Resultat från random forest som visar att variansen inte förklaras särskilt väl av modellen i fråga.*

Procentuell andel av variansen som är förklarad beräknas med hjälp av OOB-MSE (ekvation (2.13)) enligt

$$1 - \frac{\mathrm{MSE}_{OOB}}{\hat{\sigma}_y^2},$$

där $\hat{\sigma}_y^2$ beräknas med n som delare (snarare än $n-1$). [12]

Förklarad varians i randomForest() är ett mått på hur väl OOB-prediktioner förklarar variansen hos responsen i träningsmängden. Oförklarad varians beror då på slumpmässigt beteende hos data och brist på passning.

Att en sådan liten andel av variansen är förklarad med hjälp av modellen tyder på att modellen inte passar särskilt bra för denna datamängd (liknande resultat får vi för de andra två grenarna), men det kan också vara så att vi har en datamängd som inte är tillräckligt tillrättalagd för den här metoden. För att få en uppfattning om detta gör vi en simulering av data (för att eliminera eventuella beroenden och liknande i data). Vi utgår från GLM och simulerar en ny respons från en Poissonfördelning och sätter lambda-parametern lika med väntevärdet av motsvarande tariffcell som respektive observation tillhör. Sedan applicerar vi samma Random Forest-algoritm som tidigare på den nya simulerade datamängden.

```
Call:
 randomForest(formula = skfreq ~ ., data = Hull, mtry = 4,
                Type of random forest: regression
                      Number of trees: 100
No. of variables tried at each split: 4

          Mean of squared residuals: 106.6631
                    % Var explained: 1.5
```

Figur 23. *Resultat från random forest för simulerad datamängd som visar att variansen inte förklaras särskilt väl av modellen nu heller.*

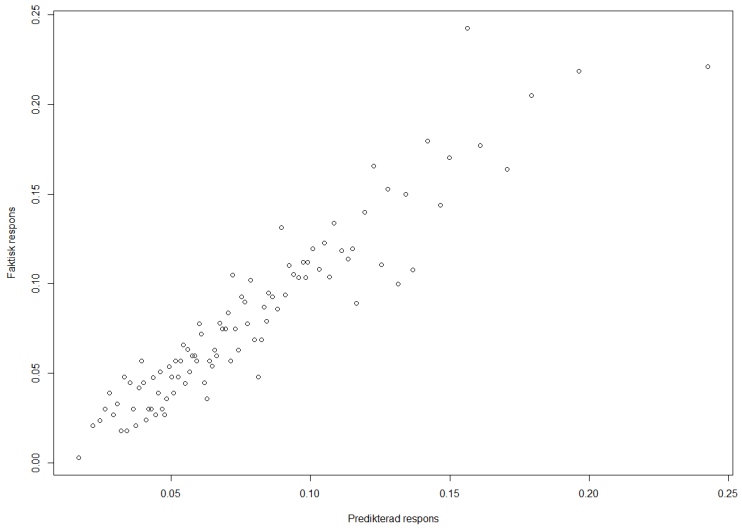
I figur 23 kan vi se att även här är det en väldigt liten andel av variansen som är förklarad, och därmed styrks teorin om att denna modell inte passar särskilt bra för denna datamängd.

5.4 Jämförelse av alla tre modellerna

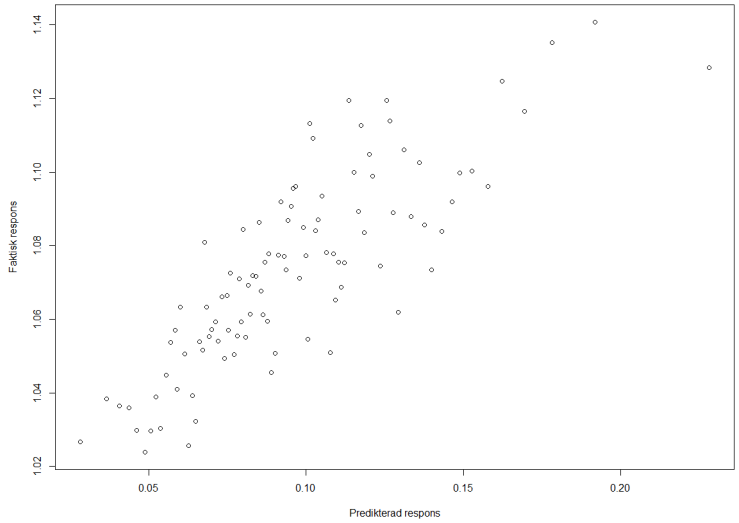
Eftersom vi har gjort alla modelleringarna på en träningsmängd per gren, så har vi också en testmängd per gren som vi kan använda för att undersöka prediktionsförmågan hos de olika modellerna.

Vi använder funktionen predict() i R för att få fram tränings- och testfelen. Processen att tilldela prediktioner till individuella observationer kallas *scoring*. Ett väldigt enkelt och lättbegripligt diagnos-verktyg för att jämföra prestationen av olika modeller är att skapa en plot av den faktiska responsvariabeln (på y-axeln) versus den predikterade responsvariabeln på testmängden (på x-axeln) för varje modell. Om en modell passar bra bör vi få ett samband som liknar en rät linje eftersom de predikterade värdena ligger nära de faktiska.

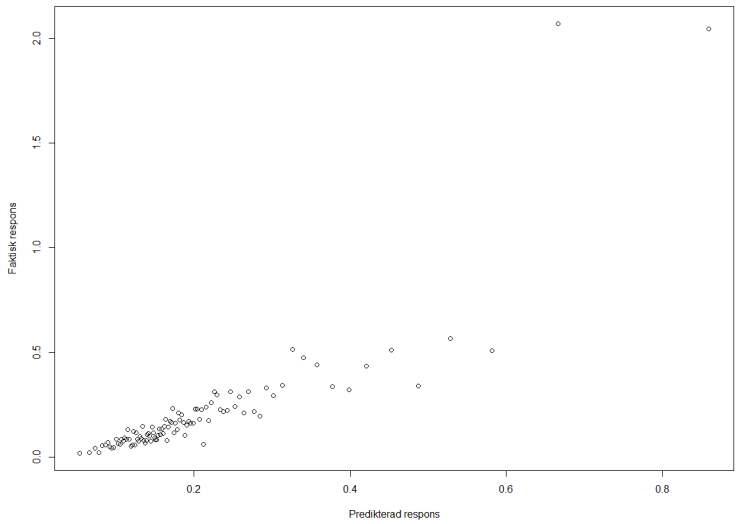
För grenen Vagnskada får vi följande grafer:



Figur 24. *Modellen GLM ger följande resultat på test-data.*



Figur 25. *Modellen Logistisk regression ger följande resultat på test-data.*



Figur 26. *Modellen Random forest ger följande resultat på test-data.*

Här kan vi se att vi mer eller mindre har ett linjärt samband för alla metoderna, men för Random Forest kan vi se en skevhet med outliers (om vi logaritmerar alla värden och plottar på nytt får vi ett linjärt samband som ser bättre ut, men eftersom vi är ute efter en jämförelse mellan modellerna så gör vi på samma sätt för alla tre) som skiljer den metoden från de övriga. Logistisk regression har ett lite svagare linjärt samband än vad GLM har. De faktiska och de predikterade värdena överensstämmer bra i GLM-fallet, inte särskilt bra alls i fallet med logistisk regression och hyfsat bra i Random Forest-fallet (bortsett från outlier-punkterna).

6. Diskussion och slutsats

6.1 Kritisk diskussion och felkällor

Eftersom de olika valideringarna av våra modeller antydde att GLM var en bättre modell än de andra två och det dessutom i jämförelsen av de tre framkom att GLM verkar fungera bäst på datamängden i fråga, är det den modell vi kan lita mest på för dessa data.

Det finns flera möjliga felkällor och vi räknar upp några här:
(1) Om man har en modell baserad på oberoende observationer är denna inte säkert tillämpbar på mätningar över tiden då det ofta finns ett beroende mellan observationer som ligger nära i tiden. Vår datamängd omfattas av flera år av försäkringsdata. Alltså kommer det att finnas fall där distinkta observationer egentligen är multipla förnyelser av samma kontrakt. Dessa observationer kommer troligen att ha korrelerade utfall. Både GLM och Logistisk regression förutsätter att observationerna är oberoende, dock så fann vi ingen nämnvärd multikollineäritet med GVIF-testet. Random forest-algoritmen bör i teorin inte ha några problem med multikollineäritet, men det kan vara ett problem för modelleraren. Betrakta två nära relaterade prediktorer, om de är bra prediktorer så kommer RF-algoritmen att använda båda ungefär lika ofta. Detta är problematiskt eftersom vi är intresserade av måttet 'Variable importance' och algoritmen kanske ger resultatet att båda dessa prediktorer är 'ganska viktiga', men om vi bara hade använt den ena så kommer den att ta över större delen av den andras vikt också och istället tolkas som en 'mycket viktig' variabel.
(2) Vi har bortsett från skadekostnaderna och därmed inte tagit hänsyn till hur allvarlig varje skada är. Men det finns indikationer på att skadefrekvensen och medelskadan kan ha ett samband (se [13]).
(3) Vi hade från början fyra bank-variabler uppdelade på fyra olika typer av bank-engagemang som vi sedan lade ihop till en enda variabel med värden 'Ja' och 'Nej', där 'Ja' betyder att en kund har ett eller fler bank-engagemang. Viss information kan ha gått förlorad här eftersom det finns nära 14000 kunder med två eller fler bankengagemang.
(4) För GLM och logistisk regression har flera av variablerna delats in i kategorier, vilket kan ha gjort modellen för grov.
(5) Data är mycket obalanserat eftersom antalet skador som finns registrerade bara utgör några procent av observationerna (i försök att motverka detta prövar vi både undersampling och elastic net, men det gav ingen nämnvärd förbättring).
(6) Det kan finnas data-kluster (men vårt resultat från PCA motsäger detta).

(7) Vi valde att dela upp datamängden i 70% träningsmängd och 30% testmängd (som är en allmänt vedertagen uppdelning), när det kunde vara så att en annan uppdelning hade fungerat bättre.
Att Random Forest inte verkar prestera så bra som väntat kan bero på data i sig. Innan vi tillämpade Random Forest-metoden på data provade vi att göra ett träd i taget med hjälp av tree()-funktionen i R, men då får vi nästan varje gång resultatet att trädet bara består av en enda rot (eller ibland en 'stubbe', dvs. ett träd med endast tre noder), vilket vanligen betyder att data har för låg varians för att bygga träd.
Det märkliga är att å ena sidan tyder grunda träd på att vi har en datamängd med låg varians och hög bias, vilket isåfall borde innebära att variablerna korrelerar mycket, men vi har ju flera andra resultat som tyder på att det inte förekommer någon större korrelation mellan variablerna. Men vi får för samma datamängd även fram att vi har överspridning, vilket brukar innebära hög varians.

6.2 Slutsats

Resultaten som vi har fått från de tre modellerna på de tre grenarna är motstridiga: Den multiplikativa GLM-modellen säger sammanlagt att bank-kunderna är 'bättre' kunder, dvs. mindre riskbenägna än de andra kunderna i samma bestånd. Den logistiska regressionsmodellen säger tvärtom att det är bank-kunderna som är 'sämre' än andra kunder. Random forest säger att Bank-variabeln inte är någon särskilt viktig variabel till att börja med.
Men enligt valideringen och jämförelsen av modellerna så verkar GLM (som nämnt) vara den mest pålitliga för just dessa data. Isåfall kan vi dra slutsatsen att bank-kunderna är mer risk-aversiva speciellt inom grenarna Trafik och Vagnskada för detta bestånd.

6.3 Vidare utveckling av arbetet

Vår slutsats till trots så bör man nog göra flera och noggrannare analyser av speciellt data innan man beslutar sig för att agera och införa ev. rabatter.
En metod som eventuellt kan vara av intresse och som lämpar sig bra för just grunda träd med hög bias men låg varians är boosting. Medan random forests bygger en ensemble av djupa oberoende träd, bygger en boosting machine (se s. 321 [8]) en ensemble av grunda och svaga successiva träd där varje träd tränas och förbättras utifrån det föregående trädet som skapats.

7. Appendix

	population	relationstal	ki_undre_gräns	ki_övre_gräns
A1	44090	1.000000e+00	NA	NA
A2	51191	1.111533e+00	1.04830693	1.1785718
A3	4197	1.367700e+00	1.18392839	1.5799972
A4	11908	1.064263e+00	0.97210782	1.1651547
B1	32946	1.000000e+00	NA	NA
B2	14991	7.635550e-01	0.70477727	0.8272347
B3	12406	7.271575e-01	0.66606522	0.7938531
B4	11629	6.923899e-01	0.63115674	0.7595637
B5	9263	6.438822e-01	0.58054393	0.7141308
B6	30151	5.430949e-01	0.50298079	0.5864083
BankJa	66275	1.000000e+00	NA	NA
BankNej	45111	1.198163e+00	1.13584797	1.2638976
C1	25	2.728184e-10	0.00000000	Inf
C2	56	2.929248e-01	0.03975247	2.1584816
C11	10179	1.016124e+00	0.90613557	1.1394623
C12	10481	1.000000e+00	NA	NA
C13	9550	8.732895e-01	0.77197990	0.9878943
C14	8217	9.105322e-01	0.80071488	1.0354109
C15	7003	8.847466e-01	0.77056457	1.0158480
C16	5847	8.800211e-01	0.75896334	1.0203882
C17	5143	9.112556e-01	0.78142972	1.0626506
C18	4094	7.416601e-01	0.61953449	0.8878597
C19	3182	7.776845e-01	0.63690198	0.9495860
C20	2115	9.679137e-01	0.77232780	1.2130301
C3	134	8.552541e-01	0.39182537	1.8667999
C21	3877	6.161660e-01	0.49162892	0.7722503
C4	2939	1.414453e+00	1.20287027	1.6632538
C5	4661	1.172993e+00	1.01416588	1.3566935
C6	4557	1.165249e+00	1.00701308	1.3483483
C7	5280	1.163986e+00	1.01596734	1.3335704
C8	6835	1.059379e+00	0.93242940	1.2036134
C9	8359	1.028261e+00	0.91040611	1.1613717
C10	8852	1.073968e+00	0.95450345	1.2083837
D1	5488	1.261712e+00	1.10984895	1.4343542
D2	5367	1.338867e+00	1.17552755	1.5249027
D3	2682	8.649919e-01	0.70957732	1.0544461
D4	5364	1.002902e+00	0.87135902	1.1543027
D5	2282	1.062794e+00	0.88133966	1.2816076
D6	3406	8.487570e-01	0.71179029	1.0120796
D7	4026	8.510930e-01	0.72496585	0.9991634
D8	1839	9.259881e-01	0.73628154	1.1645736
D9	3905	1.228055e+00	1.06049158	1.4220945
D10	1583	1.047606e+00	0.83553970	1.3134966
D11	2625	1.075013e+00	0.89632097	1.2893305
D12	3294	9.975060e-01	0.84062259	1.1836681
D13	4791	9.653760e-01	0.83236667	1.1196398
D14	5200	1.025745e+00	0.89009447	1.1820688
D15	7291	8.666082e-01	0.75690352	0.9922133
D16	5940	9.679345e-01	0.84761741	1.1053303
D17	7510	1.316801e+00	1.17379677	1.4772284
D18	15753	1.000000e+00	NA	NA
D19	14914	1.085392e+00	0.97931410	1.2029606
D20	8126	1.028892e+00	0.91142138	1.1615028
E1	312	5.207061e-01	0.29016016	0.9344315
E2	103470	1.000000e+00	NA	NA
E3	7604	6.925367e-01	0.61991417	0.7736669
F1	51303	1.202736e+00	1.12818038	1.2822180
F2	557	1.362551e+00	0.91062693	2.0387555
F3	59526	1.000000e+00	NA	NA
G1	98808	1.000000e+00	NA	NA
G2	12512	1.170737e+00	1.07570549	1.2741644
G3	15	1.766467e+00	0.36377347	8.5778849
G4	7	1.971341e-10	0.00000000	Inf
G5	44	5.152521e-01	0.07023470	3.7799657
H1	4138	1.871869e+00	1.63826523	2.1387820
H2	15	9.685051e-10	0.00000000	Inf
H3	107233	1.000000e+00	NA	NA
I1	54441	1.000000e+00	NA	NA
I2	40391	9.511060e-01	0.89776486	1.0076165
I3	11885	9.158343e-01	0.83761287	1.0013606
I4	3270	1.092189e+00	0.94100421	1.2676635
I5	1399	1.050250e+00	0.82833583	1.3316160
J1	47018	1.000000e+00	NA	NA
J2	10142	8.440905e-01	0.75943879	0.9381780
J3	41236	1.230684e+00	1.16212934	1.3032818
J4	12990	9.164738e-01	0.83583513	1.0048923
K1	47293	1.162376e+00	1.10251632	1.2254858
K2	58	2.652198e+00	0.90086678	7.8082088
K3	64035	1.000000e+00	NA	NA
L1	2254	1.141231e+00	0.94868631	1.3728537
L2	3459	8.881456e-01	0.75579690	1.0436700
L3	105673	1.000000e+00	NA	NA

Figur A1. *(Förra sidan). Tabell över alla variabler och nivåer oh deras antal observationer, duration, antalskador, relationstal och konfidensintervall. Gren: Vagnskada*

	population	relationstal	ki_undre_gräns	ki_övre_gräns
A1	44090	1.000000e+00	NA	NA
A2	51191	1.070324e+00	1.00980456	1.1344703
A3	4197	1.213519e+00	1.05357692	1.3977424
A4	11908	1.083960e+00	0.99045827	1.1862887
B1	32946	1.000000e+00	NA	NA
B2	14991	8.499664e-01	0.78477770	0.9205701
B3	12406	8.008439e-01	0.73377559	0.8740423
B4	11629	7.656338e-01	0.69804724	0.8397642
B5	9263	7.195234e-01	0.64885907	0.7978834
B6	30151	6.209759e-01	0.57518971	0.6704067
BankJa	66275	1.000000e+00	NA	NA
BankNej	45111	1.126399e+00	1.06805392	1.1879322
C1	25	1.730991e-08	0.00000000	Inf
C2	56	2.694343e-01	0.03690716	1.9669576
C11	10179	1.042632e+00	0.93033762	1.1684806
C12	10481	1.000000e+00	NA	NA
C13	9550	8.583628e-01	0.75923456	0.9704336
C14	8217	8.747621e-01	0.76983228	0.9939941
C15	7003	8.206055e-01	0.71521439	0.9415266
C16	5847	8.012788e-01	0.69150223	0.9284825
C17	5143	8.188473e-01	0.70262874	0.9542890
C18	4094	6.446903e-01	0.53876331	0.7714438
C19	3182	6.732809e-01	0.55203518	0.8211562
C20	2115	7.864783e-01	0.62831037	0.9844626
C3	134	8.565942e-01	0.39420007	1.8613737
C21	3877	4.533291e-01	0.36190589	0.5678472
C4	2939	1.391554e+00	1.18479918	1.6343893
C5	4661	1.178358e+00	1.01984770	1.3615052
C6	4557	1.172072e+00	1.01405855	1.3547075
C7	5280	1.196856e+00	1.04571120	1.3698476
C8	6835	1.106235e+00	0.97450615	1.2557705
C9	8359	1.062415e+00	0.94140342	1.1989815
C10	8852	1.090861e+00	0.97018645	1.2265451
D1	5488	1.227120e+00	1.08069003	1.3933906
D2	5367	1.303193e+00	1.14550266	1.4825921
D3	2682	8.361858e-01	0.68696546	1.0178193
D4	5364	9.889688e-01	0.85999177	1.1372892
D5	2282	1.065533e+00	0.88454554	1.2835527
D6	3406	8.251347e-01	0.69260405	0.9830252
D7	4026	8.299722e-01	0.70804119	0.9729008
D8	1839	9.069649e-01	0.72213058	1.1391088
D9	3905	1.214392e+00	1.04990447	1.4046503
D10	1583	1.049423e+00	0.83805611	1.3140981
D11	2625	1.029421e+00	0.85925606	1.2332843
D12	3294	9.750123e-01	0.82232757	1.1560465
D13	4791	9.358425e-01	0.80753729	1.0845334
D14	5200	1.019755e+00	0.88544595	1.1744371
D15	7291	8.608722e-01	0.75266415	0.9846369
D16	5940	9.644128e-01	0.84507155	1.1006075
D17	7510	1.291224e+00	1.15259572	1.4465248
D18	15753	1.000000e+00	NA	NA
D19	14914	1.090253e+00	0.98438836	1.2075026
D20	8126	9.957953e-01	0.88490816	1.1205775
E1	312	5.486850e-01	0.30676115	0.9813993
E2	103470	1.000000e+00	NA	NA
E3	7604	7.035053e-01	0.63023541	0.7852933
F1	51303	1.181982e+00	1.10993065	1.2587106
F2	557	1.143376e+00	0.77027257	1.6972027
F3	59526	1.000000e+00	NA	NA
G1	98808	1.000000e+00	NA	NA
G2	12512	1.084474e+00	0.99832497	1.1780574
G3	15	1.950655e+00	0.42101554	9.0377981
G4	7	2.170315e-08	0.00000000	Inf
G5	44	4.352790e-01	0.05989746	3.1632023
H1	4138	1.465803e+00	1.28631061	1.6703428
H2	15	3.140328e-08	0.00000000	Inf
H3	107233	1.000000e+00	NA	NA
I1	54441	1.000000e+00	NA	NA
I2	40391	9.645093e-01	0.91075860	1.0214323
I3	11885	9.358617e-01	0.85634664	1.0227599
I4	3270	1.083194e+00	0.93409588	1.2560919
I5	1399	9.914628e-01	0.78358353	1.2544910
J1	47018	1.000000e+00	NA	NA
J2	10142	8.717516e-01	0.78471838	0.9684377
J3	41236	1.243726e+00	1.17488583	1.3165987
J4	12990	9.273363e-01	0.84611316	1.0163565
K1	47293	1.158376e+00	1.09908490	1.2208650
K2	58	1.482608e+00	0.53047824	4.1436702
K3	64035	1.000000e+00	NA	NA

Figur A2. *Tabell över alla variabler (utom I som har uteslutits) och nivåer*

och deras antal observationer, duration, antalskador, relationstal och konfidensintervall. Gren: Vagnskada

	population	relationstal	ki_undre_gräns	ki_övre_gräns
A1	44090	1.000000e+00	NA	NA
A2	51191	1.069332e+00	1.00891521	1.1333660
A3	4197	1.209683e+00	1.05033788	1.3932025
A4	11908	1.087261e+00	0.99367497	1.1896616
B1	32946	1.000000e+00	NA	NA
B2	14991	8.498010e-01	0.78462649	0.9203892
B3	12406	8.003486e-01	0.73333375	0.8734876
B4	11629	7.646984e-01	0.69724695	0.8386751
B5	9263	7.182697e-01	0.64775695	0.7964583
B6	30151	6.201571e-01	0.57446550	0.6694829
BankJa	66275	1.000000e+00	NA	NA
BankNej	45111	1.128832e+00	1.07046863	1.1903764
C1	25	1.716259e-08	0.00000000	Inf
C2	56	2.704796e-01	0.03705232	1.9744844
C11	10179	1.042075e+00	0.92983474	1.1678635
C12	10481	1.000000e+00	NA	NA
C13	9550	8.587361e-01	0.75956002	0.9708617
C14	8217	8.755201e-01	0.77050058	0.9948538
C15	7003	8.222637e-01	0.71666852	0.9434175
C16	5847	8.026564e-01	0.69270722	0.9300572
C17	5143	8.213266e-01	0.70479716	0.9571227
C18	4094	6.469606e-01	0.54069302	0.7741139
C19	3182	6.745915e-01	0.55310790	0.8227575
C20	2115	7.895551e-01	0.63080107	0.9882629
C3	134	8.567415e-01	0.39431009	1.8614944
C21	3877	4.556175e-01	0.36378118	0.5706378
C4	2939	1.388655e+00	1.18261806	1.6305881
C5	4661	1.174335e+00	1.01659824	1.3565462
C6	4557	1.169250e+00	1.01173915	1.3512827
C7	5280	1.194742e+00	1.04394461	1.3673221
C8	6835	1.104871e+00	0.97333097	1.2541874
C9	8359	1.061589e+00	0.94067823	1.1980416
C10	8852	1.090417e+00	0.96978741	1.2260506
D1	5488	1.228536e+00	1.08195858	1.3949699
D2	5367	1.307046e+00	1.14898359	1.4868536
D3	2682	8.382330e-01	0.68866657	1.0202826
D4	5364	9.869775e-01	0.85827155	1.1349841
D5	2282	1.063397e+00	0.88276005	1.2809972
D6	3406	8.251386e-01	0.69261977	0.9830121
D7	4026	8.301106e-01	0.70815517	0.9730687
D8	1839	9.082064e-01	0.72309443	1.1407069
D9	3905	1.219701e+00	1.05459531	1.4106557
D10	1583	1.050672e+00	0.83903820	1.3156863
D11	2625	1.031484e+00	0.86099415	1.2357334
D12	3294	9.760886e-01	0.82323778	1.1573194
D13	4791	9.375662e-01	0.80904193	1.0865078
D14	5200	1.021109e+00	0.88661744	1.1760019
D15	7291	8.606308e-01	0.75253083	0.9842591
D16	5940	9.629937e-01	0.84387674	1.0989246
D17	7510	1.291791e+00	1.15310408	1.4471579
D18	15753	1.000000e+00	NA	NA
D19	14914	1.088674e+00	0.98301183	1.2056947
D20	8126	9.968334e-01	0.88584700	1.1217251
E1	312	5.529671e-01	0.30915377	0.9890630
E2	103470	1.000000e+00	NA	NA
E3	7604	7.037010e-01	0.63043446	0.7854823
F1	51303	1.178805e+00	1.10708551	1.2551712
F2	557	1.144484e+00	0.77097301	1.6989489
F3	59526	1.000000e+00	NA	NA
G1	98808	1.000000e+00	NA	NA
G2	12512	1.082786e+00	0.99731092	1.1755876
G3	15	1.943403e+00	0.41895771	9.0147854
G4	7	2.114216e-08	0.00000000	Inf
G5	44	4.380279e-01	0.06028566	3.1826552
H1	4138	1.471685e+00	1.29165215	1.6768115
H2	15	3.163793e-08	0.00000000	Inf
H3	107233	1.000000e+00	NA	NA
J1	47018	1.000000e+00	NA	NA
J2	10142	8.722584e-01	0.78519921	0.9689704
J3	41236	1.245768e+00	1.17696108	1.3185977
J4	12990	9.272382e-01	0.84602236	1.0162505
K1	47293	1.161323e+00	1.10212079	1.2237051
K2	58	1.473654e+00	0.52705559	4.1203529
K3	64035	1.000000e+00	NA	NA

Figur A3. *Tabell över alla variabler (utom I och G som har uteslutits) och nivåer och deras antal observationer, duration, antalskador, relationstal och konfidensintervall. Gren: Vagnskada*

	population	relationstal	ki_undre_gräns	ki_övre_gräns
A1	44090	9.084561e-01	0.8565771	0.9634772
A2	51191	1.000000e+00	NA	NA
A3	4197	1.152735e+00	1.0088740	1.3171105
A4	11908	1.039290e+00	0.9498093	1.1372004
B1	32946	1.000000e+00	NA	NA
B2	14991	8.829985e-01	0.8157103	0.9558373
B3	12406	7.559031e-01	0.6911217	0.8267566
B4	11629	7.235398e-01	0.6585303	0.7949671
B5	9263	7.252744e-01	0.6538892	0.8044528
B6	30151	6.195464e-01	0.5736828	0.6690767
BankJa	66275	1.000000e+00	NA	NA
BankNej	45111	1.130987e+00	1.0721877	1.1930103
C1	25	4.524300e-08	0.0000000	Inf
C2	56	7.372228e-01	0.2255030	2.4101568
C11	10179	1.044563e+00	0.9308663	1.1721478
C12	10481	1.000000e+00	NA	NA
C13	9550	8.246700e-01	0.7275464	0.9347590
C14	8217	9.148361e-01	0.8057387	1.0387053
C15	7003	8.494409e-01	0.7401903	0.9748167
C16	5847	8.232764e-01	0.7098854	0.9547795
C17	5143	8.254158e-01	0.7072072	0.9633828
C18	4094	6.641684e-01	0.5539896	0.7962598
C19	3182	7.408133e-01	0.6106454	0.8987284
C20	2115	6.853061e-01	0.5372939	0.8740924
C3	134	1.298781e+00	0.6947900	2.4278308
C21	3877	4.865387e-01	0.3894648	0.6078083
C4	2939	1.377127e+00	1.1680486	1.6236302
C5	4661	1.174103e+00	1.0140680	1.3593945
C6	4557	1.234690e+00	1.0678055	1.4276556
C7	5280	1.208705e+00	1.0553192	1.3843837
C8	6835	1.168914e+00	1.0298955	1.3266967
C9	8359	1.139991e+00	1.0112157	1.2851666
C10	8852	1.140285e+00	1.0138834	1.2824458
D1	5488	1.226057e+00	1.0788323	1.3933722
D2	5367	1.237126e+00	1.0843725	1.4113984
D3	2682	9.449715e-01	0.7837145	1.1394085
D4	5364	9.528790e-01	0.8271636	1.0977012
D5	2282	1.138786e+00	0.9501442	1.3648797
D6	3406	8.372428e-01	0.7044757	0.9950316
D7	4026	7.343771e-01	0.6221872	0.8667966
D8	1839	8.715962e-01	0.6926948	1.0967023
D9	3905	1.212444e+00	1.0468640	1.4042140
D10	1583	1.033038e+00	0.8193078	1.3025225
D11	2625	1.006857e+00	0.8398882	1.2070181
D12	3294	8.794373e-01	0.7341703	1.0534476
D13	4791	9.994716e-01	0.8641855	1.1559364
D14	5200	9.928743e-01	0.8608401	1.1451598
D15	7291	8.924182e-01	0.7808486	1.0199291
D16	5940	9.255215e-01	0.8090907	1.0587069
D17	7510	1.244399e+00	1.1088998	1.3964548
D18	15753	1.000000e+00	NA	NA
D19	14914	1.119759e+00	1.0113772	1.2397560
D20	8126	1.025017e+00	0.9110119	1.1532878
E1	312	7.783652e-01	0.4759543	1.2729214
E2	103470	1.000000e+00	NA	NA
E3	7604	7.265940e-01	0.6512636	0.8106378
F1	51303	1.184261e+00	1.1120264	1.2611870
F2	557	8.258595e-01	0.5126861	1.3303343
F3	59526	1.000000e+00	NA	NA
H1	4138	1.439142e+00	1.2632020	1.6395860
H2	15	7.186869e-08	0.0000000	Inf
H3	107233	1.000000e+00	NA	NA
J1	47018	1.000000e+00	NA	NA
J2	10142	8.699444e-01	0.7818364	0.9679817
J3	41236	1.250994e+00	1.1813389	1.3247555
J4	12990	9.580962e-01	0.8746067	1.0495556
K1	47293	1.161169e+00	1.1016812	1.2238682
K2	58	1.072099e+00	0.2589446	4.4387692
K3	64035	1.000000e+00	NA	NA

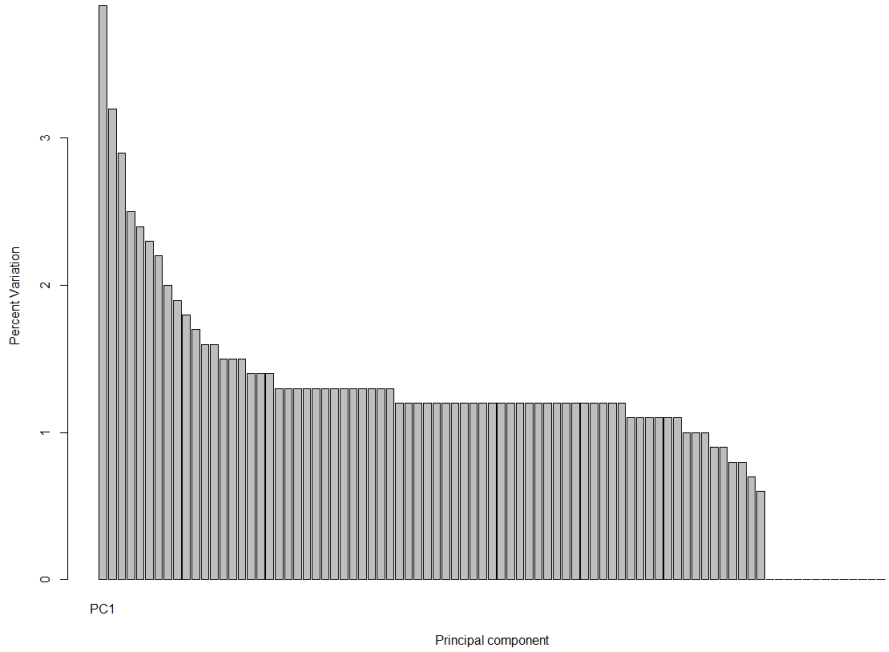
Figur A4. *Tabell över alla variabler (utom I och G som har uteslutits) och klasser efter att klass C1-C3 har lagts ihop med C4. Gren: Vagnskada.*

	population	relationstal	ki_undre_gräns	ki_övre_gräns
A1	55998	1.0000000	NA	NA
A2	55388	1.0582655	1.0016433	1.1180884
B1	32946	1.0000000	NA	NA
B2	14991	0.8727644	0.8054716	0.9456792
B3	12406	0.7957693	0.7277911	0.8700969
B4	11629	0.7587631	0.6912998	0.8328099
B5	9263	0.7351754	0.6630468	0.8151504
B6	30151	0.6334674	0.5873458	0.6832107
BankJa	66275	1.0000000	NA	NA
BankNej	45111	1.1118402	1.0538742	1.1729945
C12	59204	1.0000000	NA	NA
C13	9550	0.7911976	0.7166152	0.8735424
C14	8217	0.7974975	0.7172814	0.8866845
C15	7003	0.7043324	0.6238916	0.7951448
C16	5847	0.7588347	0.6681090	0.8618805
C17	5143	0.7460443	0.6516108	0.8541634
C18	4094	0.6138825	0.5202151	0.7244152
C19	3182	0.6264625	0.5200571	0.7546387
C20	2115	0.7728400	0.6253040	0.9551860
C21	3877	0.4303431	0.3481659	0.5319165
C4	3154	1.2422550	1.0808282	1.4277916
D1	5488	1.3140244	1.1787322	1.4648451
D2	5367	1.2083327	1.0745447	1.3587782
D7	4026	0.8412986	0.7247351	0.9766097
D9	3905	1.1861034	1.0365881	1.3571845
D15	7291	0.8180361	0.7262477	0.9214254
D17	7510	1.2577897	1.1402146	1.3874889
D20	77799	1.0000000	NA	NA
E2	103470	1.0000000	NA	NA
E3	7916	0.7407572	0.6651605	0.8249455
F1	51860	1.2265608	1.1574966	1.2997458
F3	59526	1.0000000	NA	NA
H1	4138	1.5130126	1.3385627	1.7101980
H3	107248	1.0000000	NA	NA
J1	47018	0.8224338	0.7786204	0.8687125
J2	10142	0.7120313	0.6411119	0.7907957
J3	54226	1.0000000	NA	NA
K1	47351	1.1481161	1.0893533	1.2100488
K3	64035	1.0000000	NA	NA

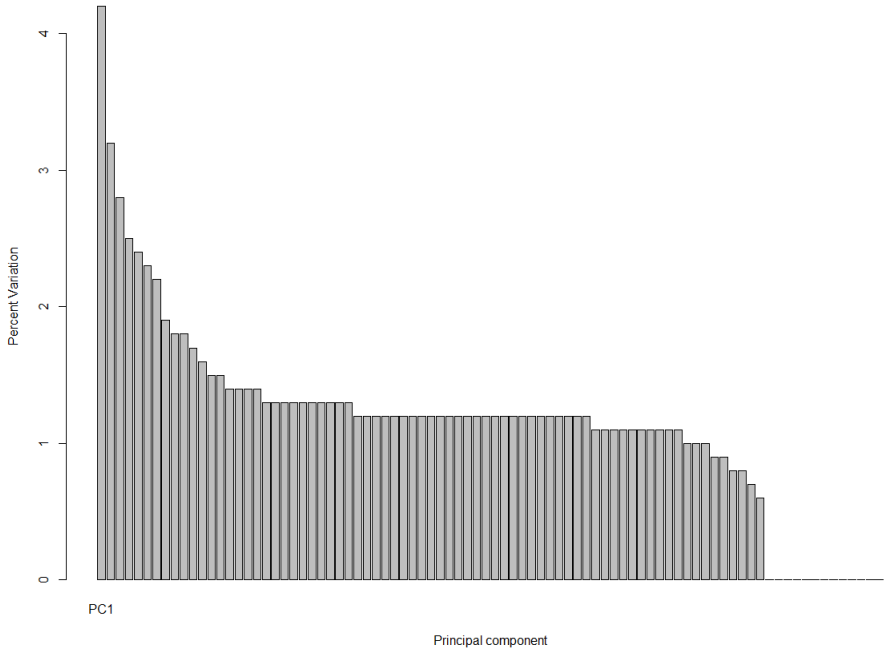
Figur A5. *Den färdiga stabila tabellen med alla klasser signifikanta. Gren: Vagnskada.*

	population	relationstal	ki_undre_gräns	ki_övre_gräns
B1	23218	1.0000000	NA	NA
B2	10472	1.1132656	1.0053885	1.2327177
B3	8608	1.3029251	1.1628791	1.4598368
B4	8125	1.3381342	1.1911703	1.5032303
B5	6472	1.5160724	1.3305750	1.7274303
B6	21075	1.6009925	1.4591819	1.7565849
BankJa	46413	1.0000000	NA	NA
BankNej	31557	0.9068678	0.8471488	0.9707965
C12	26880	1.0000000	NA	NA
C15	4913	1.1896423	1.0178989	1.3903628
C16	4121	1.2716695	1.0680757	1.5140717
C17	3610	1.4368009	1.1814097	1.7474013
C18	2869	1.6389589	1.2982270	2.0691190
C19	2263	1.2883027	1.0171526	1.6317354
C20	1490	1.5501279	1.1182339	2.1488317
C21	2685	2.3969622	1.7332919	3.3147491
C4	2249	0.6978745	0.5799730	0.8397439
C5	3279	0.7572606	0.6461180	0.8875216
C6	3178	0.7019384	0.6005783	0.8204052
C7	3672	0.7024886	0.6099129	0.8091159
C8	4726	0.7802836	0.6846161	0.8893197
C9	5889	0.8519004	0.7525677	0.9643441
C10	6146	0.7664852	0.6821056	0.8613028
D1	3837	0.8045647	0.6954504	0.9307987
D2	3737	0.8190614	0.7033284	0.9538384
D6	2380	1.4578420	1.1736941	1.8107813
D7	2819	1.2420797	1.0346809	1.4910510
D9	2713	0.7861656	0.6657754	0.9283255
D14	3617	1.2154760	1.0226474	1.4446639
D15	5070	1.2629371	1.0860316	1.4686591
D17	5244	0.8402843	0.7403332	0.9537295
D20	48553	1.0000000	NA	NA
E2	72454	1.0000000	NA	NA
E3	5516	1.3618434	1.1911628	1.5569808
F1	36200	0.7921132	0.7352168	0.8534127
F3	41770	1.0000000	NA	NA
H1	2874	0.6302692	0.5218683	0.7611868
H3	75096	1.0000000	NA	NA
J1	32914	1.1234360	1.0484984	1.2037294
J2	7050	1.2437481	1.0951798	1.4124706
J3	38006	1.0000000	NA	NA
K1	33185	0.8274075	0.7742675	0.8841946
K3	44785	1.0000000	NA	NA

Figur A6. *(förra sidan) Den färdiga stabila tabellen med alla klasser signifikanta för Logistisk regressions-modellen. Gren: Vagnskada.*



Figur A7. *Scree plot. Visar hur stor procentandel av den totala variansen som beskrivs av principal komponent nr x (x-axeln). Gren: Delkasko.*



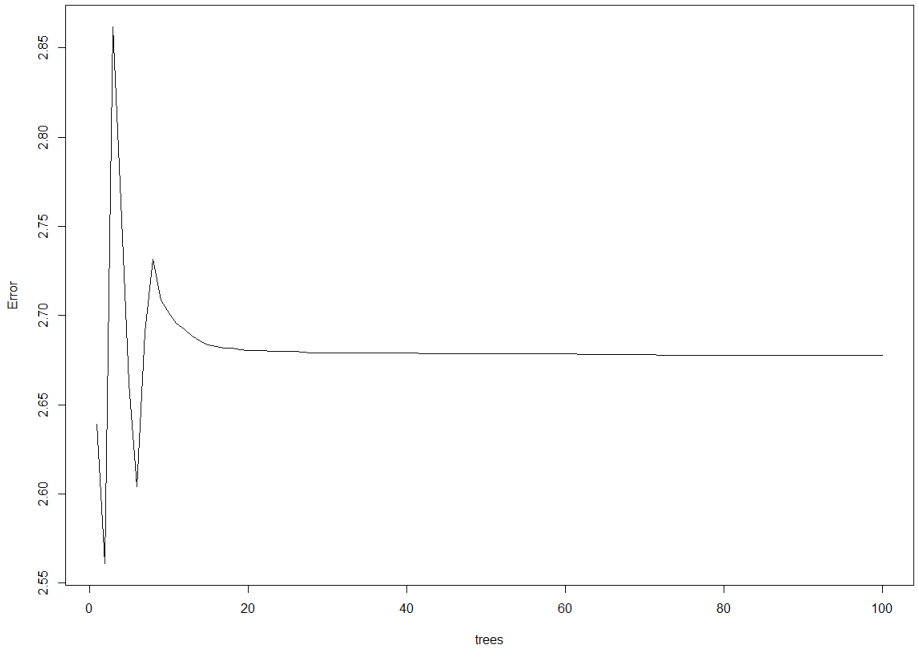
Figur A8. *Scree plot. Visar hur stor procentandel av den totala variansen som beskrivs av principal komponent nr x (x-axeln). Gren: Trafik.*

```
Call:
 randomForest(formula = skfreq ~ ., data = Casco, mtry = 4,
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 4

          Mean of squared residuals: 2.367694
                    % Var explained: 0.07
```

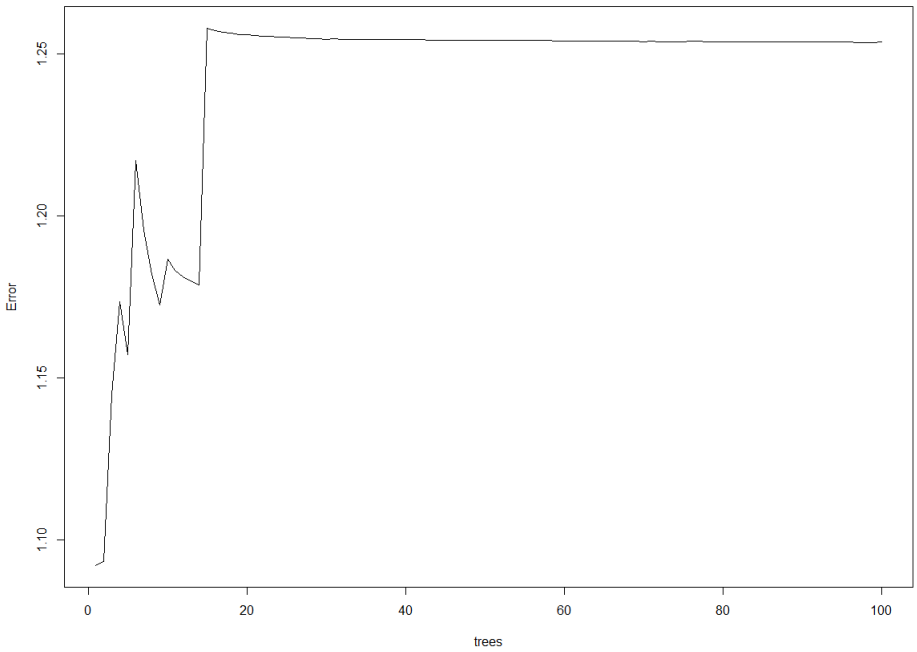
Figur A9. *Resultat från random forest som visar att variansen inte förklaras särskilt väl av modellen i fråga. Delkasko.*

```
Call:
 randomForest(formula = skfreq ~ ., data = TPL, mtry = 4,
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 4

          Mean of squared residuals: 1.125276
                    % Var explained: 0
```

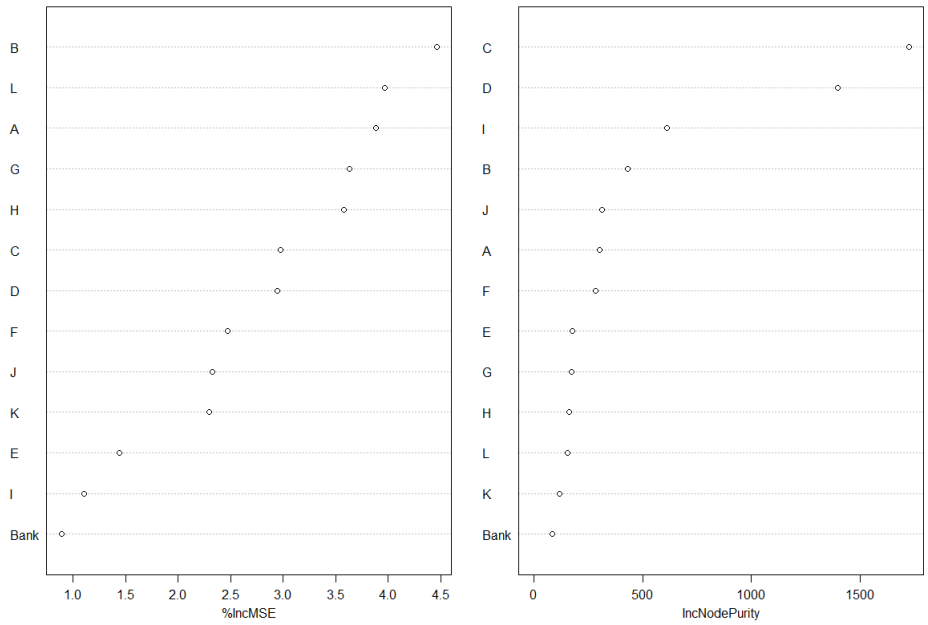
Figur A10. *Resultat från random forest som visar att variansen inte förklaras särskilt väl av modellen i fråga. Trafik.*



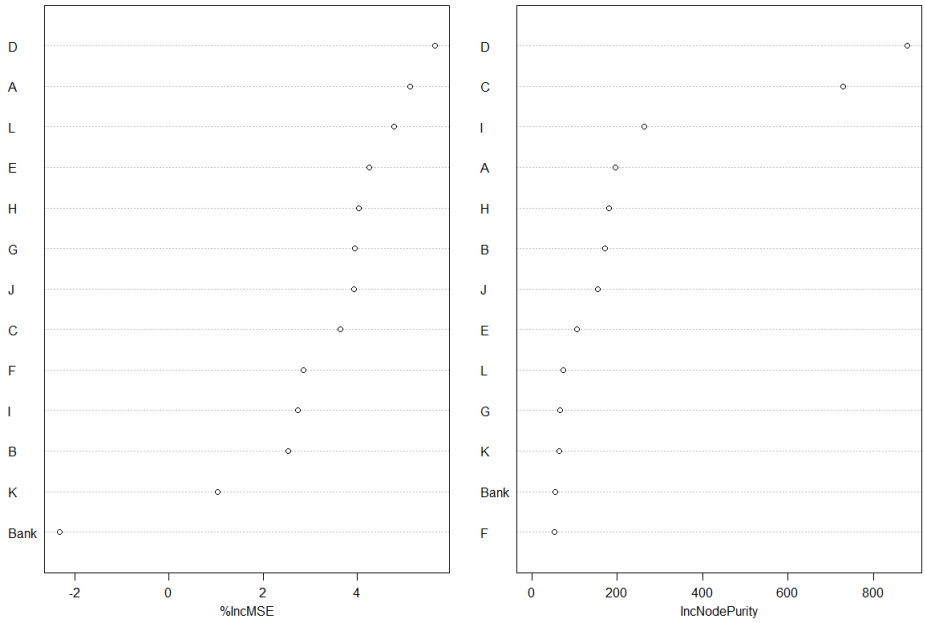
Figur A11. *Visar hur OOB-felet (y-axeln) minskar med antalet träd (x-axeln), med stoppregel nodesize=1000. Delkasko.*



Figur A12. *Visar hur OOB-felet (y-axeln) minskar med antalet träd (x-axeln), med stoppregel nodesize=1000. Trafik.*



Figur A13. *Variable importance. Delkasko.*



Figur A14. *Variable importance. Trafik.*

8. Källhänvisningar

[1] Agresti, Alan *Categorical Data Analysis. Third Edition.* 2013. Wiley.

[2] Anton, Howard & Rorres, Chris *Elementary linear algebra. Tenth edition.* 2011. Wiley.

[3] Breiman, Leo. 2001. *Random Forests.* Machine learning 45(1), 5-32.

[4] Dunn, Peter K. & Smyth, Gordon K. 1996. *Randomized Quantile Residuals.* University of Queensland.

[5] Goldburd, Mark & Khare, Anand & Tevet, Dan *Generalized Linear Models for Insurance Rating.* 2016. Casualty Actuarial Society.

[6] Held, Leonard & Bové, Daniel Sabanés *Applied Statistical Inference.* 2014. Springer.

[7] Henckaerts, Roel & Côté, Marie-Pier & Antonio, Katrien & Verbelen, Roel *Boosting insights in insurance tariff plans with tree-based machine learning methods.* 2019. `https://arxiv.org/abs/1904.10890`.

[8] James, Gareth & Witten, Daniela & Hastie, Trevor & Tibshirani, Robert *An Introduction to Statistical Learning.* 2013. Springer.

[9] Johansson, Björn *Matematiska modeller inom sakförsäkring.* 2014. Matematisk statistik. Stockholms universitet.

[10] de Jong, Piet. & Heller, Gillian Z. *Generalized Linear Models for Insurance Data.* 2008. Cambridge university press.

[11] Kaufman, Jennifer D. & Dunlap, William P. 2000. *Determining the number of factors to retain: A Windows-based FORTRAN-IMSL program for parallel analysis.* Instruments & Computers32 (3), 389-395

[12] Liaw, Andy. 2014. *Classification and Regression by RandomForest.* `https://www.researchgate.net/publication/228451484`.

[13] Oh, Rosy & Peng, Shi & Ahn, Jae Youn *Bonus-Malus premiums under the dependent frequency-severity modeling.* 2019. Scandinavian Actuarial Journal.

[14] Ohlsson, Esbjörn & Johansson, Björn *Non-Life Insurance Pricing with Generalized Linear Models.* 2010. Springer.

[15] *Företagsinformation om ICA försäkring* 2020. https://www.icaforsakring.se/om-ica-forsakring/foretagsinformation2/.

[16] ICA försäkring. 2020. *Bilförsäkring.* https://www.icaforsakring.se/bilforsakring.

[17] University of Cambridge. 2020. *Collinearity.* http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Collinearity.

[18] *Package 'randomForest'* CRAN 2018. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

[19] *StatQuest* Youtube. 2020. https://www.youtube.com/watch?v=J4WdyOWc_xQ&t=261s.