Stockholms universitet

Bedrägerianalys inom skadeförsäkring

- en modelljämförelse av logistisk regression, Random Forest och Boosting

Marika Lisinski

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2019:4
Försäkringsmatematik
Juni 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Masteruppsats **2019:4**
http://www.math.su.se/matstat

Bedrägerianalys inom skadeförsäkring
- en modelljämförelse av logistisk regression, Random Forest och Boosting

Marika Lisinski*

Juni 2019

Sammanfattning

Försäkringsbedrägerier ökar i Sverige. Bedrägerierna är svåra att upptäcka och man antar ett stort mörkertal bland dem. Målet med detta arbete är att hjälpa Dina Försäkringar AB att förbättra sina metoder för att identifiera potentiella försäkringsbedrägerier. För att göra detta tar vi fram tre olika modeller för att prediktera vilka skador som är bedrägerier, där vi använder logistisk regression, Random Forest och Boosting. Andelen kända bedrägerier är mycket liten i förhållande till totala antalet skador. Det stora mörkertalet försvårar modelleringen då vi antar att det finns bedrägerier i data som är felklassade som icke-bedrägerier. Vi förväntar oss på grund av detta en viss felmarginal av prediktionsgraden där vi antar att observerade icke-bedrägerier, klassade som bedrägerier i modellen, i själva verket är oupptäckta bedrägerier. Vi antar att 5-10% av antalet skador är bedrägerier och sätter en klassifikationsgräns utifrån detta. Det visar sig att vi upptäcker flest kända bedrägerier med en Random Forest-modell, där 98.28% av dem hittas i testdata då vi använder 90%-kvantilen som klassifikationsgräns. Med 90%-kvantilen som klassifikationsgräns klassar modellen även ett stort antal observerade icke-bedrägerier som bedrägerier. Vi betraktar dessa som potentiella oupptäckta bedrägerier, men utan möjlighet att utreda dessa vidare kan vi inte att säga hur väl modellen predikterar nya bedrägerier. Vi kan förvänta oss att träffsäkerheten för oupptäckta bedrägerier inte är lika hög som för upptäckta då modellen är anpassad efter observerade bedrägerier. Vi har antagit att upptäckta och oupptäckta bedrägerier beter sig på samma sätt då vi har klassat de potentiella bedrägerierna. Detta är troligtvis ett falskt antagande då det finns en anledning till varför de inte har upptäckts.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: marika.lisinski@live.se. Handledare: Mathias Lindholm.

Abstract

Insurance fraud is an increasing problem in Sweden. Frauds are hard to detect and it is assumed to be a large amount of hidden cases. The goal of this thesis is to help improve the methods for Dina Försäkringar AB for identifying potential frauds. We will do this by evaluating three different models for detecting frauds - logistic regression, Random Forest and Boosting.

The proportion of detected frauds in observed data is very small related to the total number of claims. The large number of undetected frauds complicates the analysis since we assume that are misclassified observations in the data. Because of this, we expect a certain error rate when predicting new frauds. We assume that claims predicted as frauds in the model, observed as non-frauds, are potential undetected frauds.

We assume that $5 - 10\%$ of the total number of claims are actual frauds, and determine the classification limit on this basis. Of the three models that we have analyzed, Random Forest detect the largest number of observed frauds. 98.28% is found in test data when using the 90%-quantile as classification limit.

With the 90%-quantille as classification limit we will also classify a large number of observed non-frauds as frauds. We regard these as potential undetected frauds, but with no possibility of investigating these further we can not tell how well the prediction is for new claims.

We can expect that the accuracy for undetected frauds is lower than for detected frauds, since the model is based on observed cases. We have assumed that detected and undetected frauds behave the same way when classifying the potential frauds. This is most likely an incorrenct assumption since there is a reason why they have not been detected.

Sammanfattning

Försäkringsbedrägerier ökar i Sverige. Bedrägerierna är svåra att upptäcka och man antar ett stort mörkertal bland dem. Målet med detta arbete är att hjälpa Dina Försäkringar AB att förbättra sina metoder för att identifiera potentiella försäkringsbedrägerier. För att göra detta tar vi fram tre olika modeller för att prediktera vilka skador som är bedrägerier, där vi använder logistisk regression, Random Forest och Boosting.

Andelen kända bedrägerier är mycket liten i förhållande till totala antalet skador. Det stora mörkertalet försvårar modelleringen då vi antar att det finns bedrägerier i data som är felklassade som icke-bedrägerier. Vi förväntar oss på grund av detta en viss felmarginal av prediktionsgraden där vi antar att observerade icke-bedrägerier, klassade som bedrägerier i modellen, i själva verket är oupptäckta bedrägerier.

Vi antar att $5-10\%$ av antalet skador är bedrägerier och sätter en klassifikationsgräns utifrån detta. Det visar sig att vi upptäcker flest kända bedrägerier med en Random Forest-modell, där 98.28% av dem hittas i testdata då vi använder 90%-kvantilen som klassifikationsgräns.

Med 90%-kvantilen som klassifikationsgräns klassar modellen även ett stort antal observerade icke-bedrägerier som bedrägerier. Vi betraktar dessa som potentiella oupptäckta bedrägerier, men utan möjlighet att utreda dessa vidare kan vi inte att säga hur väl modellen predikterar nya bedrägerier.

Vi kan förvänta oss att träffsäkerheten för oupptäckta bedrägerier inte är lika hög som för upptäckta då modellen är anpassad efter observerade bedrägerier. Vi har antagit att upptäckta och oupptäckta bedrägerier beter sig på samma sätt då vi har klassat de potentiella bedrägerierna. Detta är troligtvis ett falskt antagande då det finns en anledning till varför de inte har upptäckts.

Förord och tack

Detta arbete är skrivet som ett masterarbete om 30 hp i försäkringsmatematik vid Stockholms Universitet och i samarbete med Dina Försäkringar AB.

Ett stort tack till mina handledare Mathias Lindholm, Stockholms Universitet, och John Brandel, Dina Försäkringar AB. Tack Mathias för ett engagerat stöd och motiverande idéer under arbetets gång. Och tack John för uppdraget och för förtroendet och stödet att genomföra det.

Jag vill även tacka Sanna Kronman för goda tips och idéer samt resten av kollegiet på Dina Försäkringar AB för engagemang och inspiration.

Innehållsförteckning

Sammanfattning **3**

Förord och tack **4**

1 Bakgrund **6**

2 Data **6**
2.1 Urval av data 7
2.2 Bakgrundsvariabler 7
2.3 Felklassade icke-bedrägerier 9

3 Teori **10**
3.1 Generaliserade Linjära Modeller 10
3.1.1 Logistisk Regression 10
3.2 Beslutsträd 11
3.2.1 Bagging 13
3.2.2 Random Forest 14
3.2.3 Boosting 15
3.3 Undersampling 16
3.3.1 Justering av skattad sannolikhet 17
3.4 ROC 18
3.4.1 AUC 19

4 Modellering **20**
4.1 Fördelning 20
4.2 Balansera data 20
4.3 Variabelselektion Logistisk Regression 22
4.4 Modellering Random Forest och Boosting 23
4.5 Klassifikationsgräns 25

5 Resultat **26**

6 Diskussion **32**

7 Utveckling och förbättring **34**

Appendix **35**
A. Teori 35
B. Tabeller och resultat 36

Referenser **39**

1 Bakgrund

Försäkringsbedrägerier ökar i Sverige, och så även åtgärder mot dessa. Enligt en granskning av Svensk Försäkring (2017) betalas varje år cirka 60 miljarder kronor i skadeersättning i Sverige, utav dessa uppskattas $5 - 10\%$ vara försäkringsbedrägerier.

Bedrägerierna är svåra att upptäcka och kan uttrycka sig på olika sätt. Det kan till exempel gälla skador som inte inträffat eller felaktiga uppgifter angående ersättning. Det är främst inom motor- och boendeförsäkringar som bedrägerier upptäcks, fortsätter Svensk Försäkring, men även där antas det vara ett stort mörkertal.

Värdet av nekade försäkringsersättningar har ökat de senaste åren, men det antas fortfarande vara ett stort antal bedrägerier som inte upptäcks. Detta påverkar samtliga försäkringstagare i och med att premier måste höjas för att kompensera för felaktiga utbetalningar. Det är därför viktigt med en förebyggande och motverkande verksamhet mot försäkringsbedrägerier, vilket i allt större grad implementeras hos försäkringsbolagen.

Detta arbete skrivs som ett masterarbete i försäkringsmatematik vid Stockholms Universitet och i samarbete med Dina Försäkringar AB. Målet med arbetet är att hjälpa Dina Försäkringar AB att förbättra sina metoder för att identifiera potentiella försäkringsbedrägerirer. Detta görs genom att utvärdera olika analysmetoder på bolagets anmälda skador och kända bedrägerier.

Försäkringsspecifik information kommer att döljas på grund av sekretessskäl och vi kommer istället att fokusera på modellspecifika egenskaper och skillnader.

2 Data

Till analysen kommer vi att använda skadedata från Dina Försäkringar AB för en viss försäkringsgren under åren 2013-2018. Data består dels av ett antal kundspecifika variabler vilka anges vid teckning av en försäkring, och dels av ett antal skadespecifika variabler vilka är obligatoriska uppgifter vid skadeanmälan.

På grund av sekretesskäl kommer vi inte ange vilken försäkringsgren som undersöks, hur många observationer data består av eller specificera vilka variabler som ingår i analysen. Variabler kommer istället benämnas som *Variabel 1, Variabel 2* et cetera. Variabelspecifika axelvärden i figurer kommer att döljas av samma anledning.

2.1 Urval av data

Vi exkluderar observationer där försäkringstagaren är en juridisk person, detta på grund av att bedrägerier främst begås av privatpersoner och för att viss variabelinformation av intresse saknas för juridiska personer. Vi tar även bort så kallade nollskador, skador utan skadekostnad, då dessa inte är intressanta ur ett bedrägeriperspektiv. Efter rensning har vi kvar 60% av originaldata.

2.2 Bakgrundsvariabler

Angående vem som begår försäkringsbedrägeri säger Svensk Försäkring (2017)

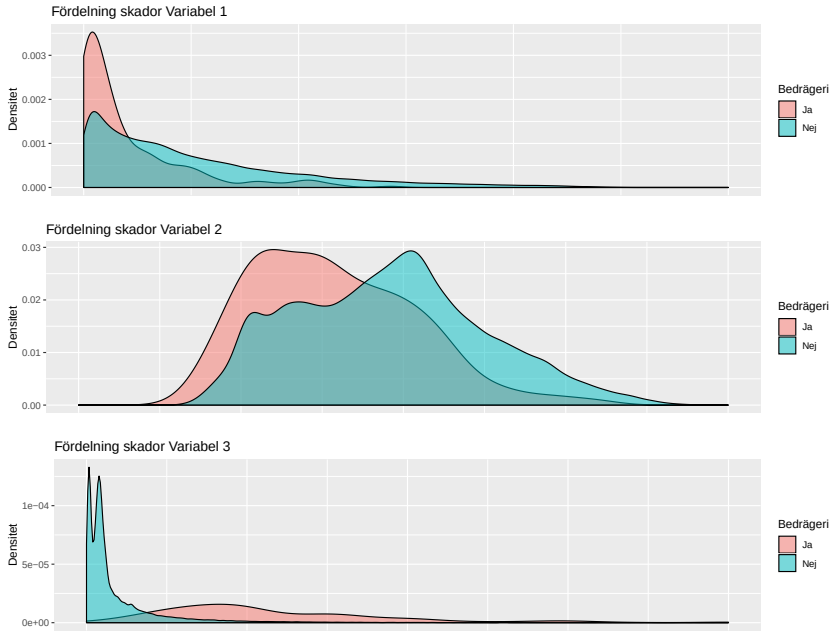
"En analys av bakgrundsvariabler till avslöjade försäkringsbedrägerier visar att det finns betydande skillnader beroende på kön, ålder och geografi när det gäller benägenheten att begå försäkringsbedrägerier."

Ovan nämnda kan alltså vara exempel på bakgrundsvariabler för att urskilja bedrägerier från resterande skador. Vi kommer i detta arbete undersöka elva stycken bakgrundsvariabler, både kategoriska och kontinuerliga. Som nämnt ovan kommer vi inte specificera vilka dessa är, utan endast benämna dem som *Variabel 1, Variabel 2,..., Variabel 11.* Dessa anges i Tabell 1 nedan.

Variabel	Variabeltyp
Variabel 1	Kontinuerlig
variabel 2	Kontinuerlig
Variabel 3	Kontinuerlig
Variabel 4	Kategorisk, 7 klasser
Variabel 5	Kategorisk, 9 klasser
Variabel 6	Kategorisk, 6 klasser
Variabel 7	Kontinuerlig
Variabel 8	Kategorisk, 2 klasser
Variabel 9	Kategorisk, 9 klasser
Variabel 10	Kategorisk, 6 klasser
Variabel 11	Kategorisk, 4 klasser

Tabell 1: Variabler med respektive variabeltyp.

I Figur 1 nedan ser vi kernel density för fördelningen av bedrägerier respektive icke-bedrägerier inom *Variabel 1,2* och *3*. Se Appendix A.3. för definition av kernel density. Observera att vi här inte tar någon hänsyn till antalet skador utan endast ser till fördelningen av dem, bedrägerierna är betydligt färre än icke-bedrägerierna vilket inte framgår av figuren.



Figur 1: Kernel density av bedrägerier respektive icke-bedrägerier inom *Variabel 1,2 och 3*.

Ovan ser vi hur bedrägerierna skiljer sig i fördelning från icke-bedrägerier inom *Variabel 1,2* och *3*. Trots tydliga skillnader observerar vi också en stor överlappning av skador, målet är att hitta kombinationer av variabler som minimerar överlappning för att i så stor grad som möjligt kunna urskilja bedrägerier från icke-bedrägerier.

2.3 Felklassade icke-bedrägerier

Vi har skadedata för försäkringsgren av intresse över perioden 2013-2018, där data är uppdaterat till och med mars 2019. I Tabell 2 nedan anges hur stor andel av totalt antal skador per år som är upptäckta bedrägerier, efter rensning av data enligt avsnitt 2.1.

Skadeår	Andel upptäckta bedrägerier av totalt antal skador
2013	0.22%
2014	0.24%
2015	0.24%
2016	0.20%
2017	0.21%
2018	0.09%

Tabell 2: Andel upptäckta bedrägerier av totalt antal skador per skadeår.

Vi tar i Tabell 2 inte hänsyn till eventuell utredningstid utan hänvisar endast skador till dess skadedatum. Andelen upptäckta bedrägerier skadeår 2018 kommer troligtvis att öka då utredningar för skador inträffade under året fortfarande pågår då data hämtas. I övrigt ser vi ingen tydlig tidstrend och antar därför att andelen bedrägerier är oberoende skadeår.

Vi observerar i Tabell 2 att upptäckta bedrägerier utgör en mycket liten del av skadorna, totalt cirka 0.19%. Enligt Svensk Försäkring (2017) utgör bedrägerier $5-10\%$ av ersättningen i skadeförsäkringsbranschen varje år, vi kommer vidare i arbetet att anta samma fördelning för antalet bedrägerier.

Med detta antagande finns det alltså strax under $5-10\%$ bedrägerier i data som inte upptäckts och alltså är felklassade som icke-bedrägerier. Detta tas hänsyn till i sådan grad att vi i modelleringen anpassar en modell som överskattar antalet bedrägerier i förhållande till observerat utfall, med grundtanken att skador modellen klassar som bedrägerier men som observerats som icke-bedrägerier är potentiella oupptäckta bedrägerier och bör utredas vidare.

3 Teori

Nedan går vi igenom den teori som sedan kommer användas i modellering av data.

3.1 Generaliserade Linjära Modeller

Agresti (2002, s. 115-117) förklarar generaliserade linjära modeller som följer. Generaliserade linjära modeller, GLM, är en generalisering av allmänna linjära modeller. Till skillnad från allmänna linjära modeller, där observationer antas vara normalfördelade, antas observationer i GLM komma från någon fördelning tillhörande den naturliga exponentialfamiljen. En stokastisk variabel Y med oberoende observationer $(y_1, ..., y_n)$ tillhör den naturliga exponentialfamiljen om dess täthets- eller sannolikhetsfördelning kan skrivas på formen

$$f_Y(y_i; \theta_i) = a(\theta_i)b(y_i)\exp\left(y_i Q(\theta_i)\right) \qquad (1)$$

där $i = 1, ..., n$, n är antalet observationer, och parameter θ_i kan variera för olika observationer. En GLM består av tre komponenter, vilka anges nedan.

- Stokastisk komponent
- Systematisk komponent
- Länkfunktion

Den stokastiska komponenten är vår responsvariabel Y med parameter θ_i och oberoende observationer $y = (y_1, ..., y_n)$, täthets- eller sannoliketsfunktionen för Y kan skrivas på samma form som (1). Parameter θ_i kan här variera för olika $i = 1, ..., n$ beroende på värden av förklaringsvariabler.

Vidare beskriver den systematiska komponenten, $\eta = (\eta_1, ..., \eta_n)$, en linjär modell av förklaringsvariabler, där

$$Q(\theta_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad \text{för } i = 1, ..., n \qquad (2)$$

och där p är antalet förklaringsvariabler, x_{ij} är värdet av förklaringsvariabel j i observation i. β_0 och β_j är okända parametrar.

Till sist har vi länkfunktionen, vilken länkar samman den stokastiska och den systematiska komponenten. Länkfunktionen är en monoton och differentierbar funktion g sådan att $g(\mu_i) = \eta_i$ där $\mu_i = E[Y_i]$.

3.1.1 Logistisk Regression

För en binär variabel Y, där $y_i \in \{0, 1\}$ för alla $i = 1, ..., n$ och $P(y_i = 1) = \pi_i$ gäller enligt Agresti (2002, s. 117) att

$$Y \sim \text{Bernoulli}(\pi)$$

där $\pi = (\pi_i, ..., \pi_n)$. Sannolikhetsfördelning kan då skrivas på formen

$$\begin{aligned} f_Y(y_i; \pi_i) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= (1-\pi_i)\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} \\ &= \exp\left(\log(1-\pi_i) + y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right) \\ &= (1-\pi_i)\exp\left(y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right) \end{aligned}$$

Vi observerar att sannolikhetsfunktionen ovan är på samma form som (1), där

$$\begin{aligned} a(\theta_i) &= 1 - \theta_i \\ b(y_i) &= 1 \\ Q(\theta_i) &= \log\left(\frac{\theta_i}{1-\theta_i}\right) = \text{logit}(\theta_i) \end{aligned}$$

Detta ger

- Stokastisk komponent: $Y \sim \text{Bernoulli}(\pi)$
- Systemastisk komponent: $\eta_i = \text{logit}(\pi_i)$
- Länkfunktion: $g(\pi_i) = \eta_i = \text{logit}(\pi_i) \quad$ där $E[Y_i] = \pi_i$

Vilket ger följande logistiska regressionsmodell

$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

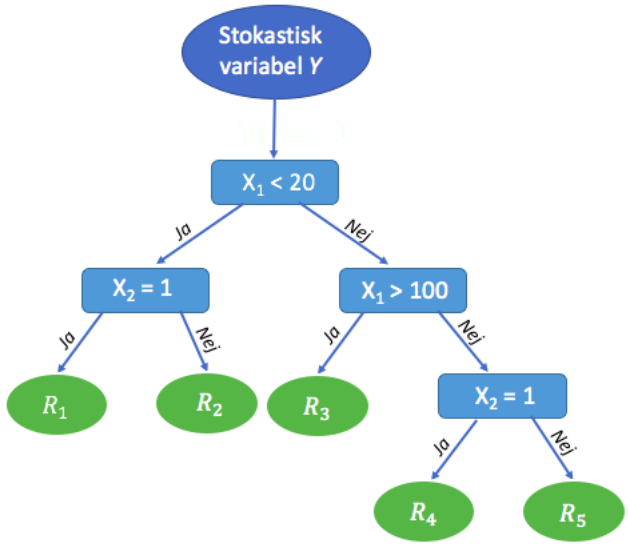
Från detta får vi enligt Sundberg (2018, s. 6) ut sannolikheten π_i genom följande

$$\pi_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}}$$

3.2 Beslutsträd

Som alternativ till GLM kommer vi att använda två olika metoder baserade på beslutsträd, Random Forest och Boosting. Innan vi går in på dessa metoder går vi här igenom vad ett beslutsträd är. Teori i följande stycke är hämtat från Hastie el al. (2009, 305-310) och James et al. (2013, 303-316).

Ett beslutsträd är uppbyggt av ett antal villkor vilka är utformade för att avgöra utfallet av stokastisk data. Villkoren delar in data i delgrupper R_j, där $j = 1, ..., m$ och m är antalet delgrupper. Se Figur 2 nedan för exempel.



Figur 2: Exempel på beslutsträd med responsvariabel Y, förklaringsvariabler X_1 och X_2, och $m = 5$ delgrupper.

I Figur 2 ovan ser vi ett exempel på hur ett beslutsträd kan vara uppbyggt. Här har vi fyra villkor vilka delar in data i $m = 5$ delgrupper beroende på observerade utfall av variabler X_1 och X_2. Utfall av delgrupp R_j bestäms av medelvärdet av observationer inom den, eller för klassifikationsträd av den vanligaste förekommande klassen i R_j.

Nedan ser vi exempel på observationer av kategorisk variabel Y tillsammans med förklaringsvariabler X_1 och X_2. Vi ser att utifrån beslutsträdet ovan hamnar $\{y_1\} \in R_1$ och $\{y_2, y_3, y_4\} \in R_3$. Utifrån detta skulle nya observationer i R_1 predikteras till $y = 0$, och i R_3 till $y = 1$ på grund av att det är den vanligast förekommande klassen.

Y	X_1	X_2
0	10	1
0	200	1
1	101	0
1	500	1

$\Longrightarrow$

R
R_1
R_3
R_3
R_3

Modellen kan skrivas som

$$f(x) = \sum_{j=1}^{m} c_j I(x \in R_j) \qquad (3)$$

där c_i är utfallet i respektive delgrupp R_j. I exemplet ovan är $c_1 = 0$ och $c_3 = 1$. Vilken förklaringsvariabel som används vid respektive förgrening väljs utifrån att man vill maximera homogenitet i delgrupperna, $R_1, ..., R_m$, där man vid varje förgrening använder den förklaringsvariabel som maximerar homogeniteten i respektive gren. Detta kan göras t.ex. genom att minimera RSS, summan av kvadratfel.

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

För klassifikationsträd kan Gini-index användas på motsvarande sätt vid val av förklaringsvariabel. Låt $\hat{p}_{mk}$ vara andelen observationer i R_j tillhörande klass k i träningsdata. Gini-index, G, används som ett variansmått över alla K klasser och fås enligt James et al. (2013, s. 312) av

$$G = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$$

För $\hat{p}_{jk}$ nära ett eller noll antar G små värden, vilket betyder att det är stor klasshomogenitet i R_j. Förgreningar i beslutsträdet väljs efter största minskning av Gini-index.

3.2.1 Bagging

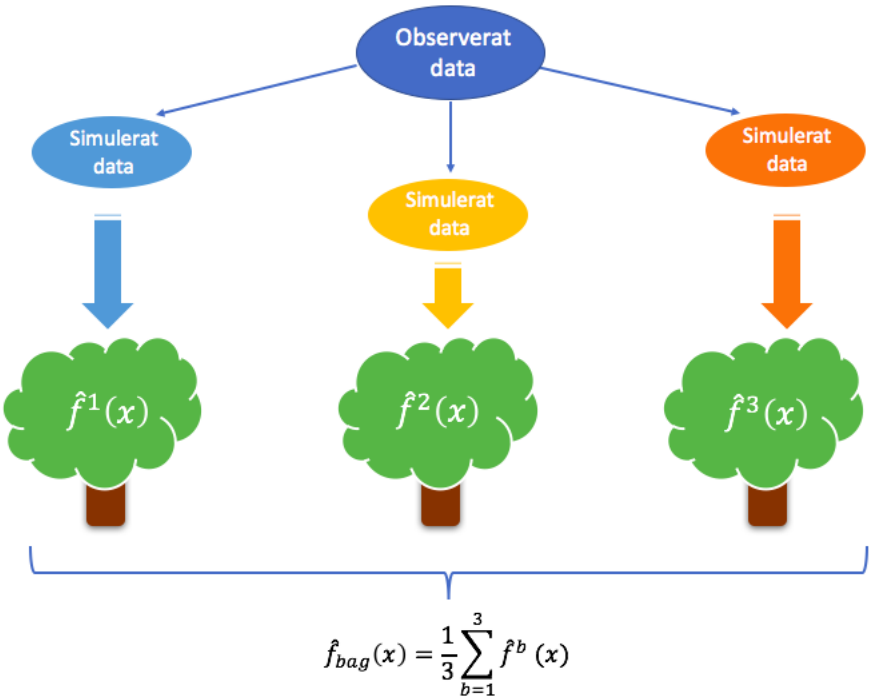
En nackdel med ett enskilt beslutsträd är att variansen kan bli stor, för att hantera detta kan man istället använda medelvärdet av flera träd, se James et al. (2003, s. 316-317). Detta baseras på principen att om variansen för n stycken oberoende stokastiska variabler $Z_1, ..., Z_n$ är lika med σ^2 så ges variansen av medelvärdet av

$$\mathrm{Var}(\bar{Z}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n}\sigma^2$$

Bagging är en metod vilken, med hjälp av bootstrap, simulerar flera dataset utifrån observerat data och sedan konstruerar ett träd per simulerat data. Resultatet ges av medelvärdet av samtliga beslutsträd, eller i klassifikationsfallet, den vanligast förekommande klassen bland samtliga träd.

Antag att $\hat{f}^b(x)$ är ett skattat beslutsträd, enligt (3), baserat på simulerat dataset b, där $b = 1, ..., B$, och B är antal bootstrapade dataset. Med bagging får vi då följande skattning

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



Figur 3: Exempel på bagging med $B = 3$ bootstrapade dataset.

I Figur 3 ovan ser vi ett exempel på hur bagging går till, där vi har använt $B = 3$ bootstrap-simulerade dataset. Tre olika beslutsträd skapas baserat på dessa simuleringar och vår modell ges av medelvärdet av dessa. I klassifikationsfallet blir utfallet den vanligast förekommande klassen, exempelvis

$$\begin{array}{l} \hat{f}^1(x) = 0 \\ \hat{f}^2(x) = 0 \\ \hat{f}^3(x) = 1 \end{array} \implies \hat{f}_{bag}(x) = 0$$

Observera att Figur 3 endast är ett exempel och att man i praktiken använder betydligt fler träd än tre.

3.2.2 Random Forest

Random Forest är likt bagging på så vis att man använder flera beslutsträd för att minska variansen hos de enskilda träden. Genom en restriktion i val av förklaringsvariabel i respektive förgrening ser man med Random Forest även till att träden inte ska korrelera med varandra, se James et al. (2013, s. 319-321).

Vilken variabel som används i respektive förgrening i ett beslutsträd väljs ut med avseende på maximerad homogenitet i delgrupperna som skapas av förgreningen. Om vår responsvariabel framför allt beror på en av förklaringsvariablerna kommer i de flesta fall den variabeln väljas som första förgreningen i respektive träd. Vi riskerar då att träden blir snarlika och korrelerande, vilket delvis tar bort den önskade effekten, minskad varians, av att använda flera träd.

I Random Forest hanterar man detta genom att begränsa antalet möjliga förklaringsvariabler vid varje förgrening. Av totalt p stycken förklaringsvariabler slumpar man fram m stycken vid respektive förgrening, där $m \leq p$. Från dessa m stycken väljs sedan den variabel som maximerar homogeniteten i delgrupperna som skapas av förgreningen på samma sätt som tidigare. Detta upprepas vid varje förgrening i trädet. Träden tvingas då till att se mer annorlunda ut och korrelationen mellan dem minskar.

Låt B vara antalet bootstrapade dataset utifrån träningsdata, precis som i bagging, och n antalet observationer i träningsdata. Modell med Random Forest fås då av algoritm nedan, se Hastie et al. (2013, s. 588)

1. Upprepa för $b = 1, ..., B$
   (a) Simulera dataset $\boldsymbol{Z^b}$ av storlek n utifrån träningsdata.
   (b) Anpassa ett beslutsträd $\hat{f}^b(x)$ på $\boldsymbol{Z^b}$ genom att upprepa följande steg för samtliga förgreningar i trädet
      i. Välj slumpvis m stycken variabler av totalt antal p förklaringsvariabler.
      ii. Välj utifrån de m variablerna vilken som ska användas till förgreningspunkt, utifrån homogeniteten i delgrupper som skapas av denna.
2. Vi har nu en samling träd, $\{f^b(x)\}_1^B$, från vilka vi får ut Random Forest-modell som
$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$$

3.2.3 Boosting

Följande teori om Boosting kommer från James et al. (2013, s. 321-323). Boosting är likt bagging en tillämpning av beslutsträd där man använder flera träd för att minska variansen. Skillnaden är att träden i boosting byggs på varandra och inte på simulerad data. Varje nytt träd baseras på information från föregående, där första trädet, $\hat{f}_1(x)$, baseras på observerad data enligt (3), och nästföljande träd baserad på residualerna från föregående.

Boosting-modellen bestäms av tre parametrar, B, λ och d, där

B: Antal träd. För många träd kan innebära en överanpassning, vilket gör modellen olämplig för prediktion. För att undvika överanpassning används korsvalidering, se Appendix A.2., för att bestämma B.

λ: Minskningsparameter vilken bestämmer hur långsamt modellen ska anpassas.

d: Antal förgreningar i varje träd. I Boosting har man gärna få förgreningar i träden för att modellen ska anpassas långsamt.

Med parametrarna ovan anpassas modellen som följer

1. Ansätt startvärden $\hat{f}(x) = 0$ och $r_i = y_i$, där y_i är observerat data av stokastisk variabel Y, $i = 1, ..., n$ och n är antalet observationer.
2. Upprepa för $b = 1, ..., B$:
   (a) Anpassa beslutsträd $\hat{f}^b(x)$ med d stycken förgreningar till data (X, r), där X är våra observerade förklaringsvariabler och r är residualer.
   (b) Uppdatera $\hat{f}(x)$ genom att addera förminskad version av $\hat{f}^b(x)$.
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$
   (c) Uppdatera residualer, r, genom att subtrahera förminskad version av $\hat{f}^b(x)$.
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$
3. Boostad modell ges av summan
   $$\hat{f}(x) = \lambda \sum_{b=1}^{B} \hat{f}^b(x)$$

3.3 Undersampling

Vi vill modellera sannolikheten för att en skada är ett bedrägeri. Data klassas i bedrägerier och icke-bedrägerier, där vi har en stor obalans mellan klasserna. Bedrägerier utgör endast 0.19% av skadorna, se Tabell 2. På grund av obalansen mellan klasserna finns det risk för att sannolikheten för bedrägeri underskattas. Detta hanterar vi med undersampling, enligt King och Zeng (2001, s. 138).

Undersampling går ut på att vi exkluderar en del av icke-bedrägerierna för att bedrägerierna ska utgöra en större del av datan. Vi antar här att icke-bedrägerierna är oberoende och väljer slumpvis vilka som exkluderas. Detta innebär en viss informationsförlust då vi inte använder all data. Ett alternativ till undersampling är så kallad oversampling, där man istället duplicerar data från den underrepresenterade klassen tills den utgör önskad andel av data. För stora dataset är ofta undersampling att föredra då de stora datamängderna kan innebära tunga körningar, varför vi i detta arbete kommer använda undersampling för att balansera data.

3.3.1 Justering av skattad sannolikhet

Då vi anpassar modeller med det balanserade datasetet måste vi ta hänsyn till det ändrade förhållandet mellan klasserna. Vi vill att de skattade sannolikheterna ska representera fördelningen i det ursprungliga, obalanserade datat.

Låt τ vara den observerade andelen bedrägerier i obalanserat data, och $\bar{y}$ motsvarande andel i balanserat data med undersampling. Justering av skattad sannolikhet ges då för logistisk regression, se King och Zeng (2001, s. 144), respektive för beslutsträd, se Pozzolo et al. (2015, s. 160-161), enligt nedan.

3.3.1.1 Logistisk regression I logistisk regression justerar vi interceptet för att ta hänsyn till balanseringen vid undersampling. Om

$$\text{logit}(\pi) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

så ges motsvarande modell med balanserat data av

$$\text{logit}(\pi^*) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \log\left(\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right)$$

För att modellen ska representera obalanserat data behöver vi alltså endast justera interceptet enligt

$$\hat{\beta}_0^* = \hat{\beta}_0 - \log\left(\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right)$$

där $\hat{\beta}_0$ är skattat intercept med balanserat data. Skattad sannolikhet för att observation i är ett bedrägeri ges då av

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0^* + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}}}{1 + e^{\hat{\beta}_0^* + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}}}$$

3.3.1.2 Beslutsträd Med beslutsträdbaserade metoder skattas sannolikheten för bedrägeri av hur många observationer av respektive klass som hamnar i varje delgrupp. Då vi balanserar om data påverkar det förhållandet av klasser i delgrupperna, vilket vi tar hänsyn till genom att se till apriori och posterior sannolikhet för en observation att tillhöra respektive klass.

Låt (X, Y) vara vårt observerade, obalanserade data, där Y är vår binära responsvariabel och X utgör våra förklaringsvariabler. I (X, Y) utgör observationer av $y = 1$ endast 0.19% av data. Vi exkluderar observationer av $y = 0$ för att balansera klasserna. Detta bildar balanserat dataset $(X, Y)'$, för vilket det gäller att $(X, Y)' \in (X, Y)$.

Vi ansätter sedan binär variabel s, där

$$s_i = \begin{cases} 1 & \text{om observation } (x,y)_i \text{ ingår i } (X,Y)' \\ 0 & \text{annars} \end{cases}$$

för $i = 1, ..., n$.

Vi vet att samtliga observationer av $y = 1$ finns med i balanserat data $(X,Y)'$ och vi får enligt Bayes formel för betingade sannolikheter

$$\begin{aligned} P(y=1|x,s=1) &= \frac{P(s=1|y=1)P(y=1|x)}{P(s=1|y=1)P(y=1|x) + P(s=1|y=0)P(y=0|x)} \\ &= \frac{P(y=1|x)}{P(y=1|x) + P(s=1|y=0)P(y=0|x)} \end{aligned}$$

Låt $\beta = P(s=1|y=0)$ vara sannolikheten för ett icke-bedrägeri att väljas från obalanserat originaldata till balanserat dataset. Låt vidare $p = P(y=1|x)$ vara sannolikheten för att skada y är ett bedrägeri i originaldata och
$p_s = P(y=1|x,s=1)$ motsvarande sannolikhet i balanserat data, detta ger

$$\begin{aligned} p_s &= \frac{p}{p + \beta(1-p)} \\ \Longrightarrow p &= \frac{\beta p_s}{\beta p_s - p_s + 1} \end{aligned}$$

Här ges β av önskad andel bedrägerier i balanserat data och p_s är våra skattade sannolikheter från modellen anpassad på balanserat data.

3.4 ROC

ROC, reciever operating characteristics, används för att visualisera hur väl en binär variabel predikteras, se Fawcett (2005, s. 861-862). Klassningen vid prediktion kan delas upp enligt Tabell 3 nedan.

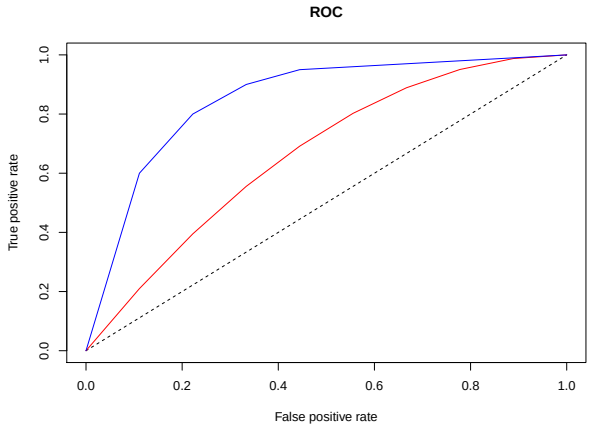
	Observerade $Y=1$	Observerade $Y=0$
Predikterade $Y=1$	Sanna $Y=1$	Falska $Y=1$
Predikterade $Y=0$	Falska $Y=0$	Sanna $Y=0$

Tabell 3: Prediktionsklassning

ROC-kurvan fås genom att plotta "True positive rate" mot "False positive rate", där dessa ges av Tabell 3 enligt

$$\text{True positive rate} = \frac{\text{Sanna } Y = 1}{\text{Observerade } Y = 1}$$
$$\text{False positive rate} = \frac{\text{Falska } Y = 1}{\text{Observerade } Y = 0}$$

I ROC-kurvan ser vi sedan hur dessa mått utvecklas för alla möjliga klassificeringsgränser, där en diagonal linje betyder att klassificering sker slumpvis. Se Figur 4 nedan för exempel.



Figur 4: Exempel på ROC kurvor.

I Figur 4 ovan ser vi exempel på två ROC-kurvor. Den streckade diagonalen representerar slumpvis klassificering. Om vi jämför den blåa och den röda ROC-kurvan ser vi att den blåa sträcker sig mer mot det vänstra hörnet och att den röda ligger närmre diagonalen. Detta betyder att den blåa kurvan representerar en klassificering med högre "True positive rate" i förhållande till "False positive rate", medan den röda kurvan motsvarar större andel felklassificering.

3.4.1 AUC

AUC, area under an ROC curve, anger arean under ROC-kurvan som andel av hela figuren, Fawcett (2005, 868). Vi ser i Figur 4 att arean är större under den blå kurvan än under den röda, det gäller att desto större AUC desto större andel korrekta klassningar ger modellen. För en diagonal ROC-kurva, det vill säga slumpvis klassificering, får vi AUC= 0.5.

4 Modellering

I följande avsnitt tillämpar vi teorin från avsnitt 4 på vår observerade data. För att kunna utvärdera och jämföra prediktionsförmågan hos de olika modellerna delar vi upp data i 80% träningsdata respektive 20% testdata. Träningsdatat använder vi till modelleringen för att sedan prediktera på testdata.

4.1 Fördelning

Låt Y_i representera varje enskild skada där $i = 1, ..., n$. Det gäller att $Y_i \in \{0, 1\}$ för alla $i = 1, ..., n$, där

$$Y_i = \begin{cases} 1 & \text{om skada } i \text{ är ett bedrägeri} \\ 0 & \text{om skada } i \text{ är ett icke-bedrägeri} \end{cases}$$

Vi antar att varje skada Y_i är ett bedrägeri med respektive sannolikhet π_i för $i = 1, ..., n$. Det gäller då att

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

vilket vi känner igen som en logistisk regressionsmodell från avsnitt 3.1.1.

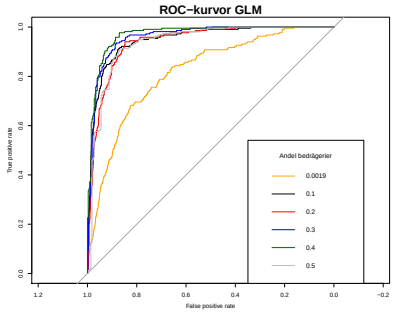
4.2 Balansera data

Originaldata består av 0.19% bedrägerier, denna låga andel gör dem svåra att upptäcka och sannolikheten för att en skada är ett bedrägeri kan underskattas. Vi balanserar om träningsdata genom undersampling, exkluderar slumpvis valda icke-bedrägerier, så att bedrägerierna utgör önskad andel av data.

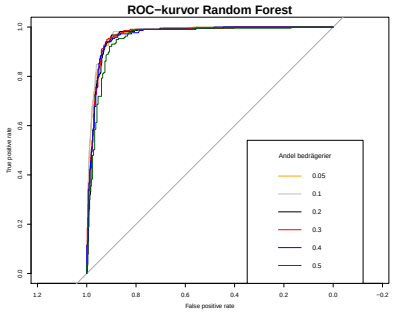
Vilken som är önskad andel testar vi oss fram till genom att jämföra ROC och AUC för modeller av olika balanserade data. Utifrån träningsdatat tar vi fram fem olika balanserade dataset, bestående av $10\%, 20\%, 30\%, 40\%$ respektive 50% bedrägerier. Ett stort antal icke-bedrägerier måste exkluderas för att bedrägerierna ska utgöra önskad andel av data, för respektive balanserat dataset används endast $1.9\%, 0.9\%, 0.6\%, 0.5\%$ och 0.4% av ursprunglig träningsdata.

För vardera balanserat dataset tar vi fram en modell med logistisk regression, Random Forest respektive Boosting, för vilka vi sedan jämför ROC och AUC dem sinsemellan. I dessa modeller görs ingen variabelselektion utan samtliga elva förklaringsvariabler används.

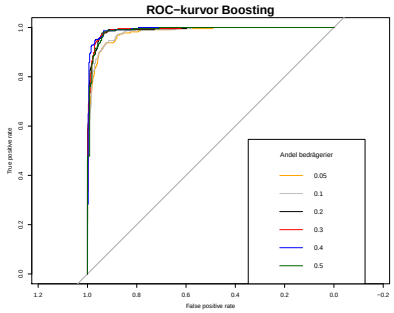
I Figur 5 nedan ser vi ROC-kurvor för respektive modell. För logistisk regression jämför vi med originalträningsdata där bedrägerierna utgör 0.19%, För Random Forest och Boosting blir det för tunga körningar på grund av stort data, varför de modellerna jämförs med en lägsta andel om 5% bedrägerier.



(a) Logistisk regression



(b) Random Forest



(c) Boosting

Figur 5: ROC-kurvor för modeller av olika balanserade data med logistisk regression, Random Forest och Boosting.

I Figur 5 ser vi ROC-kurvor för respektive balanserat dataset och modell. I Figur 5(a) ser vi en tydlig förbättring av AUC med samtliga balanserade dataset jämfört med originaldata för logistisk regression. För Random Forest och Boosting är ROC-kurvorna snarlika, se Figur 5(b) och 5(c). I Tabell 4 nedan ser vi hur AUC skiljer sig mellan respektive modell.

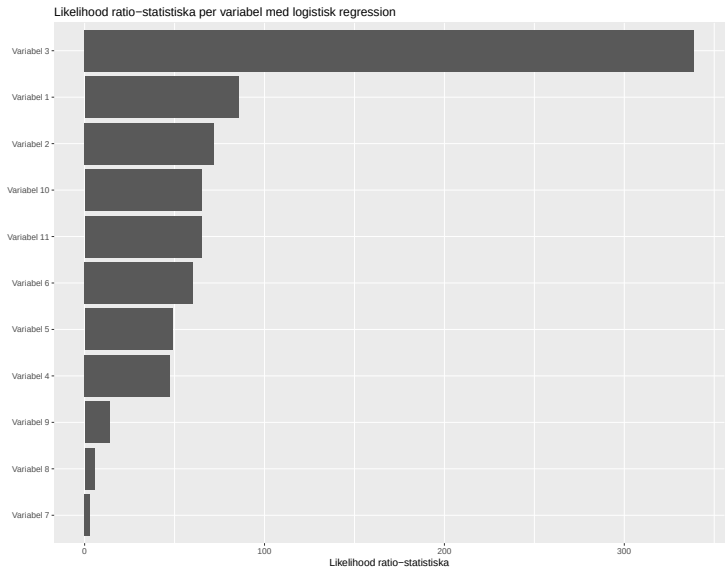
Andel kända bedrägerier i data	Logistisk regression	Random Forest	Boosting
0.0019	0.817	-	-
0.05	-	0.973	0.980
0.10	0.949	**0.974**	0.985
0.20	0.940	0.966	0.983
0.30	0.960	0.968	0.989
0.40	**0.970**	0.966	**0.992**
0.50	0.937	0.956	0.989

Tabell 4: AUC per andel bedrägerier i balanserat data.

I Tabell 4 ovan ser vi AUC-värdet för respektive modell och balanserat träningsdata. För logistisk regression och Boosting blir AUC högst med balanserat data till 40% bedrägerier, och Random Forest med 10%. Dessa andelar kommer användas i vidare modellering.

4.3 Variabelselektion Logistisk Regression

Vi har elva potentiella förklaringsvariabler i data, se Tabell 1. För att välja ut vilka variabler vi vill inkludera i den logistiska regressionsmodellen börjar vi med att undersöka förklaringsvariablerna var för sig. Vi tar fram elva stycken univariata logistiska regressionsmodeller för respektive variabel för att se de enskilda variablernas påverkan på sannolikheten för att en skada är ett bedrägeri. I Figur 6 nedan ser vi respektive likelihood ratio-statistika för vardera variabel. Se Appendix A.1. för definition av likelihood ratio-statistika.



Figur 6: Likelihood ratio-statistika för vardera förklaringsvariabel i univariat logistisk regression.

I Figur 6 ser vi att likelihood ratio-statistikan för *Variabel 3* är betydligt högre än för övriga variabler. Vidare observerar vi låga värden för *Variabel 7, Variabel 8* och *Variabel 9*, medan övriga variabler är relativt jämna. Utifrån detta anpassar vi en multivariat logistisk regressionsmodell, där *Variabel 7, 8* och *9* inte bidrar signifikant till respektive univariata modell, varför de inte inkluderas i den multivariata logistiska regressionsmodellen.

Vi har då åtta förklarande variabler i modellen. På denna modell utför vi bakåt eliminering, vilket går ut på att exkludera den förklaringsvariabel som bidrar minst till modellen, tills att samtliga variabler bidrar signifikant. I vår slutgiltiga logistiska regressionsmodell har vi då kvar förklaringsvariabler *Variabel 1, 2, 3, 10* och *11*. Se Appendix B Tabell 12 för fullständig lista över koefficientskattningar.

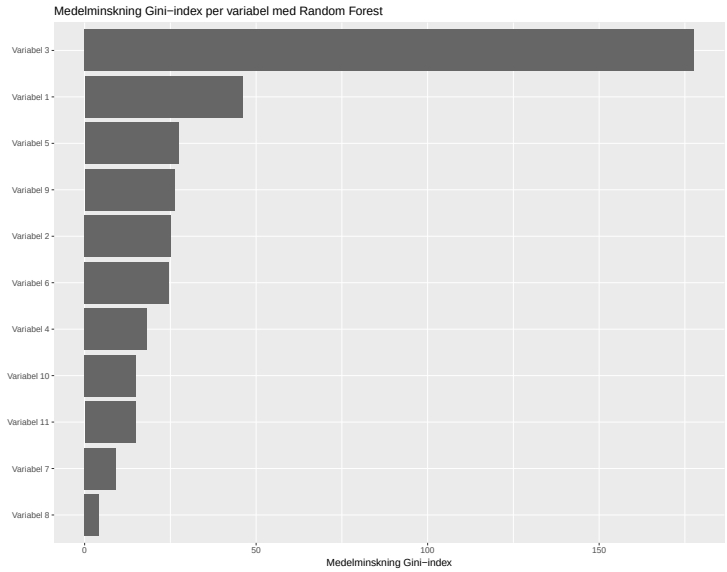
4.4 Modellering Random Forest och Boosting

Random Forest och Boosting är till skillnad från logistisk regression inte lika begränsande vad gäller val av variabler, här använder vi samtliga elva variabler i respektive modell. Random Forest- och Boosting-modellerna styrs av vissa parameterantaganden, vilka vi går igenom nedan.

Parameterantaganden Random Forest

I avsnitt 3.2.2 går vi igenom hur m av p variabler väljs ut vid varje förgrening i Random Forest. Här är $p = 11$ och m sätts till standardvärde, enligt James et al. (2013, s. 319), nämligen $m \approx \sqrt{p}$. Vi använder $B = 500$ träd i modelleringen.

I Figur 7 nedan ser vi medelminskningen av Gini-index för respektive förklaringsvariabel, vilket kan ses som ett mått på betydelsen av respektive variabel.

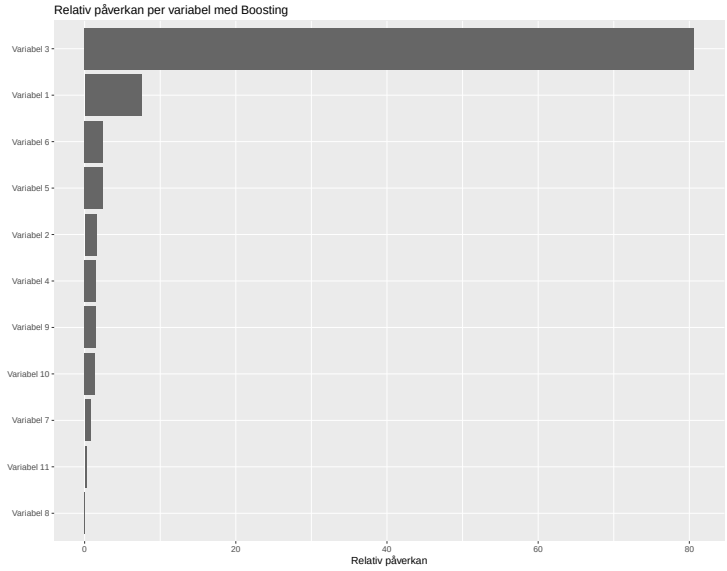


Figur 7: Medelminskning av Gini-index per variabel med Random Forest

Parameterantaganden Boosting

I Boosting-modeller använder man överlag fler träd än i Random Forest då mindre träd används. Här använder vi $B = 1052$ träd vilket fås av korsvalidering med $K = 5$ delar. Minskningsparameter sätts till standardvärde $\lambda = 0.01$, även antal förgreningar d sätts till standardvärde $d = 1$.

I Figur 8 nedan ser vi den relativa betydelsen av respektive variabel i modellen.



Figur 8: Relativ påverkan per variabel med Boosting.

4.5 Klassifikationsgräns

För att avgöra om en skada är ett bedrägeri eller inte använder vi en klassifikationsgräns. Skattas sannolikheten för bedrägeri till högre än klassifikationsgränsen säger vi att skadan är ett bedrägeri, och är sannolikheten lägre är det ett icke-bedrägeri enligt modellen.

Vi har antagit att $5-10\%$ av alla skador är bedrägerier, varav endast 0.19% i vår data är upptäckta. Vi vill därför överskatta antalet bedrägerier med modellen jämfört med faktiskt utfall i data. Utifrån detta väljer vi en klassifikationsgräns sådan att önskad andel observationer klassas som bedrägerier.

I Tabell 5 anges 90%- respektive 95%-kvantilen för den skattade sannolikheten för bedrägeri i träningsdata med respektive modell. Med 90%-kvantilen som klassifikationsgräns klassas 10% av skadorna i träningsdata som bedrägerier, respektive 5% med 95%-kvantilen som gräns. Då kända bedrägerier endast utgör 0.19% av data kommer majoriteten av de skador som klassas som bedrägerier enligt modellen vara klassade som icke-bedrägerier i data.

Modell	90%-kvantil	95%-kvantil
Logistisk regression	0.0023	0.0074
Random Forest	0.0026	0.0077
Boosting	0.0024	0.0085

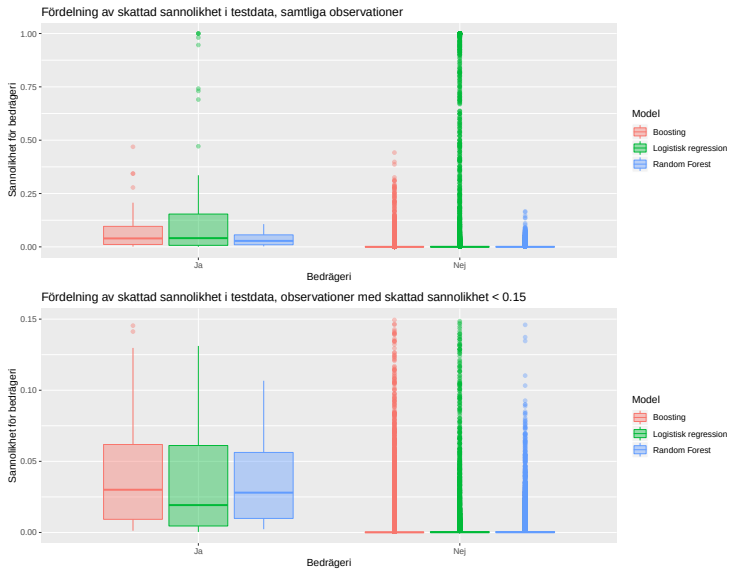
Tabell 5: 90%- och 95%-kvantil av skattad sannolikhet för bedrägeri i träningsdata med respektive modell, avrundat till fyra decimaler.

Vi observerar i Tabell 5 att kvantilerna ovan ges av mycket låga sannolikheter. Detta behöver inte påverka prediktionsförmågan, utan träffsäkerheten avgörs av hur separerade skattningarna är för bedrägerier och icke-bedrägerier.

5 Resultat

Vi har nu tre modeller för att skatta sannolikheten att en skada är ett bedrägeri. Vi observerar i Figur 6, 7 och 8 att *Variabel 3* bidrar mest till samtliga modeller. I den logistiska regressionsmodellen såg vi i Figur 6 att *Variabel 7, 8* och *9* inte bidrog signifikant till modellen. Vi observerar dock i Figur 7 och 8 att Random Forest- och Boosting-modellen verkar fånga upp en större betydelse av *Variabel 9*. Då detta arbete inte fokuserar på de specifika förklaringsvariablerna kommer vi inte gå inte mer på dem, utan fokusera på prediktionsförmågan hos de olika modellerna.

Vi vill jämföra hur väl modellerna predikterar bedrägerier bland nya skador. För att göra detta applicerar vi respektive modell på testdatat, som vi exkluderade från modelleringen, för att se hur väl de observerade bedrägerierna urskiljs från resterande skador. Nedan ser vi i boxplottar hur de skattade sannolikheterna för bedrägeri i testdata fördelas för respektive modell.



Figur 9: Boxplottar över skattad sannolikhet för bedrägeri i testdata för respektive modell. I den övre bilden är samtliga observationer inkluderade, i den nedre har vi begränsat $Y-$axeln till sannolikheter upp till 0.15.

I Figur 9 ovan ser vi hur den skattade sannolikheten för bedrägeri fördelas i testdata mellan icke-bedrägerier och kända bedrägerier med respektive modell. Vi får överlag låga sannolikheter för samtliga skador, men kan ändå observera att sannolikheten för bedrägeri tenderar att vara högre bland kända bedrägerier med samtliga modeller. I den nedre delen av figuren har vi begränsat $Y-$axeln till sannolikheter upp till 0.15 för att tydligare visualisera skillnader mellan bedrägerier och icke-bedrägerier, detta exkluderar 1.6% av observationerna med logistisk regression, 0.0% med Random Forest och 0.2% med Boosting. Se Tabell 6 nedan för en sammanställning av skattningarna i Figur 9.

Modell	Bedrägeri	Medelvärde	Median	Std.avv	Min	Max
Logistisk regression	Ja	0.1795	0.0409	0.3020	0.0004	1.0000
	Nej	0.0121	0.0002	0.0919	0.0000	1.0000
Random Forest	Ja	0.0352	0.0280	0.0285	0.0023	0.1067
	Nej	0.0016	0.0001	0.0060	0.0000	0.1665
Boosting	Ja	0.0759	0.0396	0.0962	0.0012	0.4690
	Nej	0.0024	0.0001	0.0141	0.0000	0.4419

Tabell 6: Sammanställande statistik över skattade sannolikheter för bedrägeri i testdata med logistisk regression, Random Forest och Boosting.

Vi observerar i Tabell 6 en högre variation i den skattade sannolikheten för bedrägeri med logistisk regression jämfört med de övriga modellerna, vilket stämmer överens med vad vi sett i Figur 9. Lägst variation fås med Random Forest. Se Appendix Tabell 13 och 14 för standardkvantiler inom respektive modell och klassning.

Beroende på vår klassifikationsgräns spelar det ingen roll att vi överlag får väldigt låga sannolikheter. Vi undersöker träffsäkerheten av kända bedrägerier i testdata genom att se hur stor andel av dem som klassas som bedrägerier i respektive modell då vi använder klassifikationsgränserna från Tabell 5.

Modell	Andel kända bedrägerier över 90%-kvantilen	Andel kända bedrägerier över 95%-kvantilen
Logistisk regression	84.48%	74.14%
Random Forest	98.28%	82.76%
Boosting	94.83%	81.03%

Tabell 7: Andel kända bedrägerier i testdata över 90%- respektive 95%-kvantilen i vardera modell.

I Tabell 7 ser vi hur stor andel av de kända bedrägerierna i testdata som klassas som bedrägerier med respektive klassifikationsgräns och modell. Vi ser till exempel att med 90%-kvantilen som klassifikationsgräns i den logistiska regressionsmodellen klassas 84.48% av observerade kända bedrägerier även som bedrägerier enligt modellen, motsvarande med Random Forest klassar 98.28% kända bedrägerier som bedrägerier enligt modellen.

Vi är även intresserade av de observerade icke-bedrägerier som klassas som bedrägerier enligt modellerna, då dessa är potentiella oupptäckta bedrägerier. I Tabell 8, 9 och 10 nedan ser vi respektive prediktionsklassning för modellerna där vi har använt 90%-kvantilerna från Tabell 5 som klassifikationsgräns. Här är den övre högra cellen i vardera matris potentiella oupptäckta bedrägerier enligt respektive modell. Angivna procentsatser i vardera cell anger andelen skador inom respektive klass. Se Appendix B, Tabell 15, 16 och 17 för motsvarande matriser med 95%-kvantiler som klassifikationsgräns.

	Observerade bedrägerier	Observerade icke-bedrägerier
Predikterade bedrägerier	49 (84.48%)	2 905 (10.06%)
Predikterade icke-bedrägerier	9 (15.52%)	25 977 (89.94%)

Tabell 8: Prediktionsklassning med logistisk regression, klassifikationsgräns 90%-kvantil.

	Observerade bedrägerier	Observerade icke-bedrägerier
Predikterade bedrägerier	57 (98.28%)	2 953 (10.22%)
Predikterade icke-bedrägerier	1 (1.72%)	25 929 (89.78%)

Tabell 9: Prediktionsklassning med Random Forest, klassifikationsgräns 90%-kvantil.

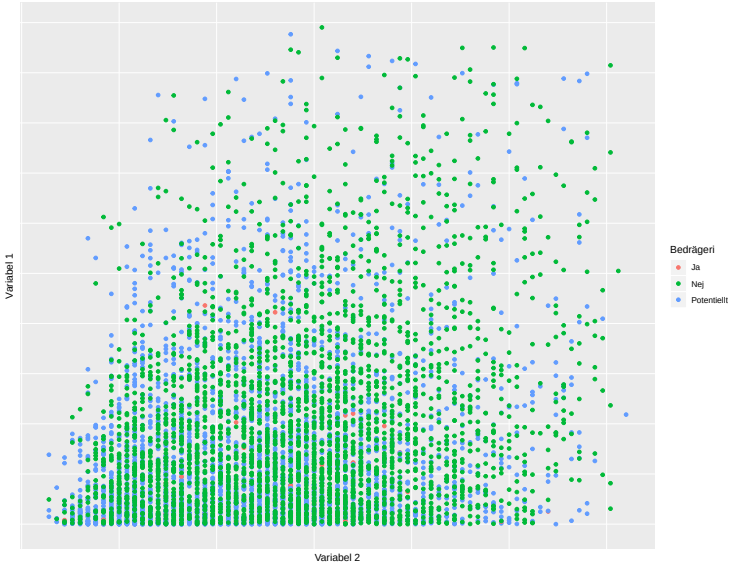
	Observerade bedrägerier	Observerade icke-bedrägerier
Predikterade bedrägerier	55 (94.83%)	2 943 (10.19%)
Predikterade icke-bedrägerier	3 (5.17%)	25 939 (89.81%)

Tabell 10: Prediktionsklassning med Boosting, klassifikationsgräns 90%-kvantil.

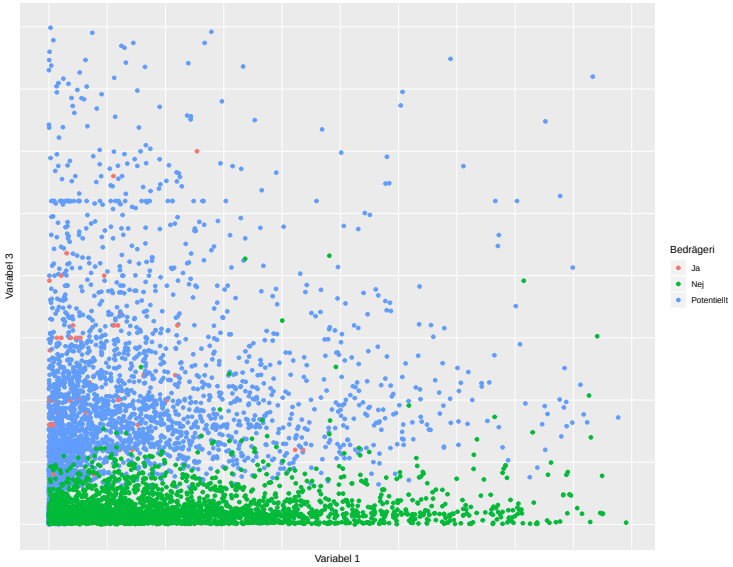
Ovan har vi använt 90%-kvantilen utifrån träningsdata som klassifikationsgräns med syfte att klassa 10% av skadorna som bedrägerier. Då vi applicerar denna på testdata ser vi i Tabell 8 att totalt 10.21% av skadorna klassas som bedrägerier med den logistiska regressionsmodellen, motsvarande 10.40% med Random Forest och 10.36% med Boosting, se Tabell 9 och 10.

Då vi använder 95%-kvantilen som klassifikationsgräns avser vi istället att klassa 5% av skadorna som bedrägerier. I testdata klassas totalt 5.21%, 5.30% respektive 5.29% av skadorna som bedrägerier med logistisk regression, Random Forest respektive Boosting, se Appendix B, Tabell 15, 16 och 17.

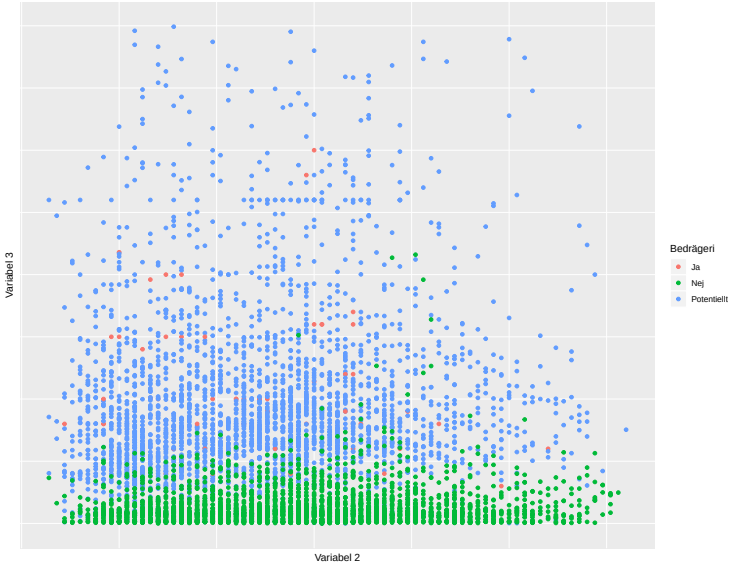
I Figur 10, 11 och 12 nedan ser vi förhållandet mellan *Variabel 1, 2* och *3* för olika klassningar av bedrägeri. Här är potentiella bedrägerier skattade enligt Random Forest med 90%-kvantilen som klassifikationsgräns. Vi har slumpvis valt ut en delmängd av icke-bedrägerier, motsvarande antalet potentiella bedrägerier, för att lättare kunna visualisera möjliga skillnader mellan grupperna.



Figur 10: Scatterplot av *Variabel 1* och *Variabel 2*, med ett slumpvis valt urval av icke-bedrägerier och skattade potentiella bedrägerier med Random Forest.



Figur 11: Scatterplot av *Variabel 1* och *Variabel 3*, med ett slumpvis valt urval av icke-bedrägerier och skattade potentiella bedrägerier med Random Forest.



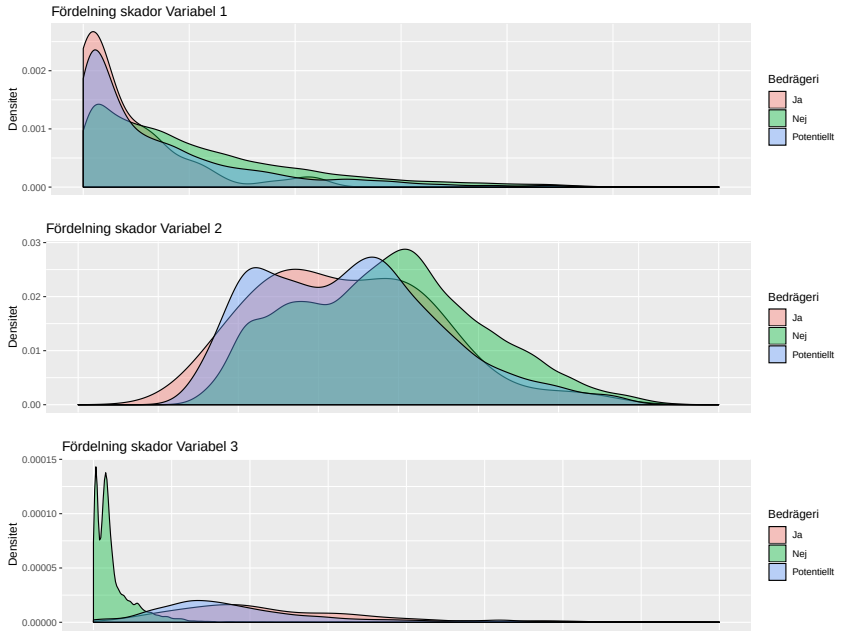
Figur 12: Scatterplot av *Variabel 1* och *Variabel 2*, med ett slumpvis valt urval av icke-bedrägerier och skattade potentiella bedrägerier med Random Forest.

Vi observerar i Figur 10, 11 och 12 ovan att det inte verkar finnas någon korrelation mellan *Variabel 1, 2* och *3*. Vi ser tydliga skillnader mellan potentiella bedrägerier och icke-bedrägerier för *Variabel 3* i Figur 11 och 12, vilket stämmer överens med att det är den mest inflytelserika variabeln i modellen.

6 Diskussion

Vi har med tre olika metoder försökt urskilja bedrägerier från icke-bedrägerier, där det har visat sig att vi får högst prediktionsgrad med Random Forest. Med 90%-kvantilen som klassifikationsgräns hittar vi 98.28% av observerade bedrägerier i testdata. Majoriteten av skador över klassifikationsgränsen är observerade icke-bedrägerier, vilket är väntat då vi antar att många icke-bedrägerier är felklassade. Vi betraktar dessa som potentiella oupptäckta bedrägerier.

Den stora mängden potentiella bedrägerier skulle innebära ett omfattande utredningsarbete för att säkerställa om de är bedrägerier eller inte. Utan möjlighet att genomföra detta går det inte att avgöra hur väl modellen predikterar nya bedrägerier. I Figur 13 nedan ser vi fördelning av skador motsvarande Figur 1 tillsammans med potentiella bedrägerier enligt modellen Random Forest med 90%-kvantilen som klassifikationsgräns, se Figur 14 för motsvarande Figur med 95%-kvantilen som klassifikationsgräns.



Figur 13: Kernel density över fördelningen av bedrägerier, icke-bedrägerier och potentiella bedrägerier inom Variabel 1,2 och 3 med 90%-kvantil som klassifikationgräns i Random Forest.

Ovan ser vi hur de potentiella bedrägerierna fördelas för *Variabel 1, 2* och *3* jämfört med icke-bedrägerier och kända bedrägerier. Vi ser att de potentiella bedrägeriena påminner i fördelning om kända bedrägerier. Vi har anpassat modellen efter upptäckta bedrägerier, inte oupptäckta, men använder modellen för att hitta båda typer. Vi antar då att upptäckta och oupptäckta bedrägerier är oberoende, vilket troligtvis är ett falskt antagande då det finns en anledning till varför vissa upptäcks och andra inte.

Antar vi samma träffsäkerhet för de potentiella bedrägerierna som för kända bedrägerier, se Tabell 7, så hittar modellen med respektive klassifikationsgräns antal oupptäckta bedrägerier i testdata enligt Tabell 11 nedan.

Klassifikationsgräns	Antal oupptäckta bedrägerier
90%−kvantil	2 902
95%−kvantil	1 229

Tabell 11: Antal oupptäckta bedrägerier i testdata enligt Random Forest-modell för respektive klassifikationsgräns.

Skulle det efter utredning visa sig att dessa faktiskt är bedrägerier finns det en stor vinst att hämta. Det är dock ett för stort antal skador för att rimligtvis kunna utreda samtliga. Antalet i Tabell 11 är även mest troligt överskattat då vi har antagit samma träffsäkerhet som för kända bedrägerier, vilka modellen är anpassad efter.

7 Utveckling och förbättring

I detta arbete har vi undersökt bedrägerierna inom en viss försäkringsgren, vilken i sin tur består av ett antal olika risker. För att minska variation i data skulle man kunna göra en finare indelning och exkludering utifrån dessa risker. Vi har valt att inte göra det på grund av att det finns så få kända bedrägerier i data från början, och en större exkludering skulle även innebära färre bedrägerier att undersöka. Utvecklingsmöjlighet finns, då fler bedrägerier utreds och upptäcks, att dela upp analysen på olika risker.

Många av variablerna vi har undersökt är försäkringsspecifika, och har inte att göra med själva skadan. De skadespecifika variabler som finns med i analysen är sådana som är obligatoriska uppgifter vid skadeanmälan, vilket är en relativt begränsad information. En utökad obligatorisk informationsplikt vid skadeanmälan skulle kunna förbättra analysen. Som tillägg till de numeriska och kategoriska variabler vi har undersökt i detta arbete skulle man även kunna undersöka skriftvariabler, till exempel anteckningar och noteringar från skadehanterare, för att ytterligare förbättra analysen.

Enligt Svensk Försäkring (2017) består $5-10\%$ av ersättningen i skadeförsäkring av bedrägerier. Vi har i detta arbete antagit samma andel för antalet skador. Detta är troligtvis överskattat då skadekostnaden tenderar att vara högre för bedrägerier än för icke-bedrägerier. En bakgrundsanalys av antalet bedrägerier skulle behövas för att motivera antalsantaganden.

Vi exkluderar en stor del observerade icke-bedrägerier då vi balanserar om träningsdata för modellering. För logistisk regression och Boosting har vi balanserat data så att kända bedrägerier utgör 40%, motsvarande 10% för Random Forest. Det betyder att vi endast använder 0.5% respektive 1.9% av originalträningsdata. För att vi inte ska förlora någon information i och med detta ska det gälla ett oberoende inom icke-bedrägerierna. Dock har vi antagit att $5-10\%$ av icke-bedrägerierna är felklassade bedrägerier. Detta är problematiskt då det säger emot antagandet om oberoende icke-bedrägerier vilket betyder att ombalansering av data kan innebära en viss informationsförlust. För att kontrollera effekten av detta kan man utföra upprepad undersampling, där vi på nytt exkluderar slumvis valda icke-bedrägerier, och jämföra resultaten av dessa. Skulle resultaten ändras beroende på vilka observationer som utesluts betyder det att undersampling inte är en lämplig metod.

Appendix

A. Teori

A.1. Likelihood ratio-test

Teori om likelihood ratio-test kommer från Held och Sabanés Bové (2014, s. 105-106). Likelihood ratio-test kan användas vid hypotestest för att testa nollhypotesen

$$H_0 : \theta = \theta_0$$

Vi har en likelihood ratio-statistika, W, som under nollhypotesen är chitvåfördelad med en frihetsgrad.

$$W \sim \chi^2(1)$$

W får vi genom andra gradens Tayolorutveckling av loglikelihoodfunktionen enligt

$$l(\theta) \approx l(\hat{\theta}_{ML}) - \frac{1}{2} l(\hat{\theta}_{ML})(\theta - \hat{\theta}_{ML})^2$$

$$\left\{ \frac{L(\hat{\theta}_{ML})}{L(\theta)} \right\} = 2\{l(\hat{\theta}_{ML}) - l(\theta)\} \approx l(\hat{\theta}_{ML})(\theta - \hat{\theta}_{ML})^2$$

Där $\hat{\theta}_{ML}$ är maximum likelihood-skattningen av θ. Likelihood ratio-statistikan ges då av

$$W = 2 \left\{ \frac{L(\hat{\theta}_{ML})}{L(\theta)} \right\}$$

A.2. Korsvalidering

Korsvalidering används som ett mått av prediktionsfel. K-delad korsvalidering går ut på att man delar upp data i K stycken lika delar. För del k, där $k \in [1, K]$, använder vi resterande $K - 1$ delar för att anpassa modellen, vilken sedan används för att prediktera på k, se Hastie et al. (2009, s. 241-243). Detta görs för samtliga $k = 1, ..., K$ delar, och korsvalideringen, CV, ges av summan av prediktionsfel från respektive modell. Låt

$$\mathbb{1}(y_i \in \text{del } k) = \begin{cases} 1 \text{ om observation } y_i \text{ ingår i del } k \\ 0 \text{ annars} \end{cases}$$

Då ges CV av

$$CV = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} L(y_i, \hat{f}^{-k}(x_i)) \mathbb{1}(y_i \in \text{del } k)$$

där n är antal observationer i data, och $\hat{f}^{-k}$ är skattad modell exklusive del k.

A.3. Kernel Density

För en stokastisk variabel X, med observationer $x_1, ..., x_n$, använder vi en kernel density funktion, med gaussiansk kärna (kernel), som en skattning av fördelningsfunktionen, $f_X(x)$. Denna ges av Hastie et al. (2009, s. 208-209) enligt nedan

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} \phi_\sigma(x - x_i)$$

där ϕ_σ är den gaussianska fördelningsfunktionen med väntevärde noll och standardavvikelse σ.

B. Tabeller och resultat

Koefficient	Estimat	Std.avv	p-värde
Intercept (ojusterat)	−0.8077	0.5782	0.1625
Variabel 1	−0.0026	0.0005	0.0000
Variabel 2	−0.0436	0.0123	0.0004
Variabel 3	0.0001	0.0000	0.0000
Variabel 10 klass 1	0.00000		
Variabel 10 klass 2	0.7385	0.3787	0.0511
Variabel 10 klass 3	1.2540	0.4836	0.0095
Variabel 10 klass 4	1.9640	0.7310	0.0072
Variabel 10 klass 5	0.0157	0.7766	0.9838
Variabel 10 klass 6	1.1660	0.7159	0.1033
Variabel 11 klass 1	0.0000		
Variabel 11 klass 2	0.4545	0.5192	0.3814
Variabel 11 klass 3	−0.1151	0.5412	0.8316
Variabel 11 klass 4	1.1210	0.4854	0.0209

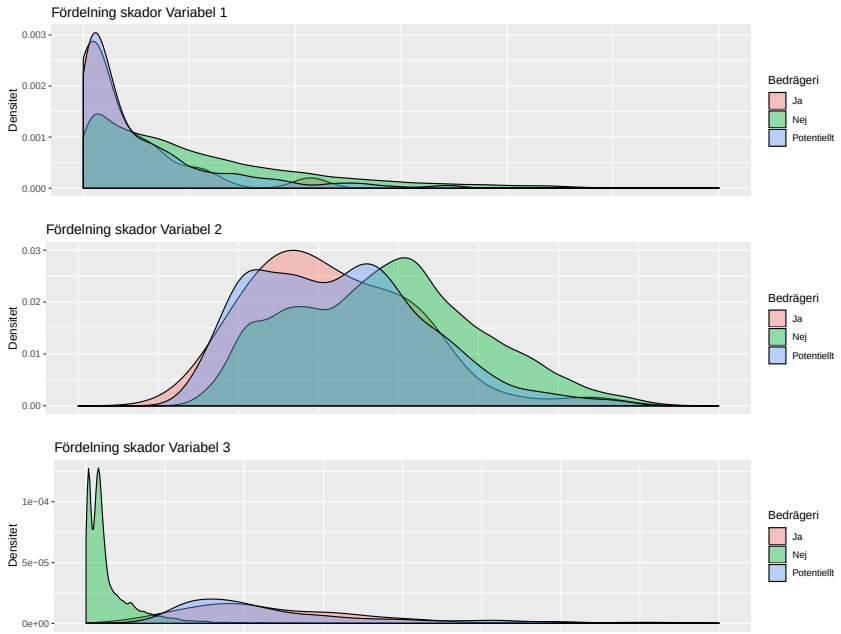
Tabell 12: Koefficientskattningar (avrundat 4 decimaler) i slutgiltig logistisk regressionsmodell.

Kvantil	Logistisk regression	Random Forest	Boosting
0%	0.00038	0.00230	0.00118
25%	0.00716	0.00987	0.01215
50%	0.04089	0.02799	0.03588
75%	0.15377	0.05627	0.09152
100%	1.00000	0.10667	0.45012

Tabell 13: Standardkvantiler för den skattade sannolikheten för bedrägeri bland observerade bedrägerier i testdata.

Kvantil	Logistisk regression	Random Forest	Boosting
0%	0.00000	0.00000	0.00000
25%	0.00005	0.00003	0.00003
50%	0.00016	0.00014	0.00007
75%	0.00055	0.00059	0.00027
100%	1.00000	0.16653	0.46417

Tabell 14: Standardkvantiler för den skattade sannolikheten för bedrägeri bland observerade icke-bedrägerier i testdata.



Figur 14: Kernel density över fördelningen av bedrägerier, icke-bedrägerier och potentiella bedrägerier inom Variabel 1,2 och 3 med 95%-kvantil som klassifikationgräns i Random Forest.

	Observerade bedrägerier	Observerade icke-bedrägerier
Predikterade bedrägerier	43 (74.14)%	1 464 (5.07%)
Predikterade icke-bedrägerier	15 (25.86%)	27 418 (94.93%)

Tabell 15: Prediktionsklassning med logistisk regression, klassifikationsgräns 95%-kvantil.

	Observerade bedrägerier	Observerade icke-bedrägerier
Predikterade bedrägerier	48 (82.76%)	1 485 (5.14%)
Predikterade icke-bedrägerier	10 (17.24%)	27 397 (94.86%)

Tabell 16: Prediktionsklassning med Random Forest, klassifikationsgräns 95%-kvantil.

	Observerade bedrägerier	Observerade icke-bedrägerier
Predikterade bedrägerier	47 (81.03%)	1 483 (5.13%)
Predikterade icke-bedrägerier	11 (18.97%)	27 399 (94.87%)

Tabell 17: Prediktionsklassning med Boosting, klassifikationsgräns 95%-kvantil.

Referenser

[1] http://www.svenskforsakring.se/globalassets/rapporter/forsakringsbedragerier/forsakringsbedragerier-i-sverige-2017.pdf 2019-02-13

[2] http://www.svenskforsakring.se/aktuellt/nyheter/2018/forsakringsbedragerierna-okar/ 2019-02-13

[3] Alan Agresti, 2002 *Categorical Data Analysis 2nd edition*, John Wiley & Sons

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009 *The Elements of Statistical Learning 2nd edition* Springer

[5] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013 *An Introduction to Statistical Learning with applications in R* Springer

[6] Rolf Sundberg, 2018 *Statistical Modelling by Exponential Families* Mathematical Statistics, Stockholm University

[7] Tom Fawcett, 2005 *An introduction to ROC analysis*, Institute for the Study of Learning and Expertise, Palo Alto

[8] Gary King, Langche Zeng, 2001 *Logistic Regression in Rare Events Data*, George Washington University

[9] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnsin, Gianluca Bontempi, 2015 *Calibrating Probability with Undersampling for Unbalanced Classification* University of Brussels

[10] Leonhard Held, Daniel Sabanés Bové, 2014. *Applied Statistical Inference* Springer