Integration of Traditional and Telematics data for Efficient Insurance Claims Prediction

Hashan Peiris

SIMON FRASER UNIVERSITY

Spring 2023

Declaration of Committee

Name:	Hashan Peiris
Degree:	Master of Science
Thesis title:	Integration of Traditional and Telematics data for Efficient Insurance Claims Prediction
Committee:	Chair:
	Himchan Jeong
	Gary Parker
	Joan Hu

Abstract

Keywords:

Dedication

Acknowledgements

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Modeling Claim Counts

1.2 Telematics in Insurance

1.2.1 Uses of Telematics Data

1.2.2 Challenges

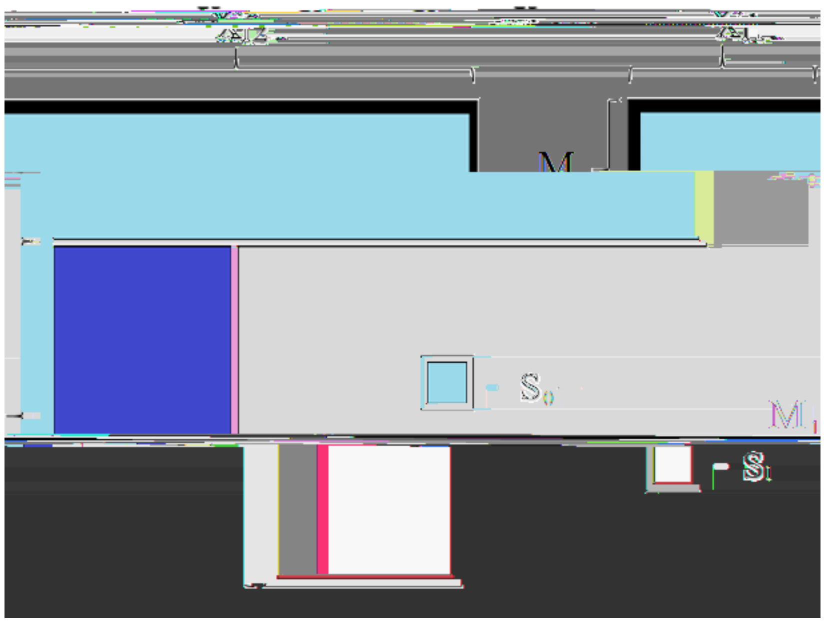
1.3 Motivation

1.4 Summary

Chapter 2

Data structure and problem description

- $i1$ … th … 0 … 1
- $i2$ … th … 0
- … $i = (i1, i2)$



… 0 1 $i1$ … $i2$

2.2 Problem Description

… $i2$ … th … 0 …

[illegible]

- [illegible] 0 [illegible]
- [illegible] $i2$ [illegible] 0 [illegible]

[illegible]

Chapter 3

Methodology

(,) = (, β)

(·) β

β

$$\sum_{i-1} \; _i - (, {}_i\beta) \; (, {}_i\beta) = 0$$

(·)

i $i2$ 1

3.2 Proposed Method

i

3.2.1 Estimation of Parameters

$i = (\ , i1\ , i2)$ $i2$

β

$$(\beta; \ ) = \ - \ \exp(\ \beta)$$

1 x_{i2}

$$\sum_{i \in S_0} {}_i[1, {}_{1i} \cdots, {}_{Li}] = \sum_{i=1}^{M} [1, {}_{1i} \cdots, {}_{Li}]$$

${}_{li} = {}_l({}_{i1} \; {}_i)$ th $[1, {}_{1i} \cdots, {}_{Li}] = [1 \; x_{i1} \; {}_i \; {}_i * x_{i1}]$ $= 2* \; +1$ x_{i1}

$$\begin{aligned}
\sum_{i \in S_0} {}_i \; (\beta; \, {}_i \; {}_i) &= \sum_{i=1}^{M} {}_i \; {}_i \; (\beta; \, {}_i \; {}_i) + \sum_{i=1}^{M} (1 - {}_i \; {}_i) \sum_{k=0}^{L} {}_k \; {}_{ki} \\
&= \sum_{i=1}^{M} {}_i \; (\beta; \, {}_i \; {}_i) + (1 - {}_i) \sum_{k=0}^{L} {}_k \; {}_{ki} \\
&\quad + \sum_{i=1}^{M} {}_i({}_i - 1) \; (\beta; \, {}_i \; {}_i) - \sum_{k=0}^{L} {}_k \; {}_{ki}
\end{aligned}$$

$\alpha = ({}_0 \; {}_1 \; \cdots \; {}_L)$ $\hat{\alpha}$

$$\sum_{i=1}^{M} {}_i({}_i - 1) \; (\beta; \, {}_i \; {}_i) - \sum_{k=0}^{L} \hat{}_k \; {}_{ki} = 0$$

$$\sum_{i \in S_0} {}_i \; (\beta; \, {}_i \; {}_i) = \sum_{i=1}^{M} {}_i \; (\beta; \, {}_i \; {}_i) + (1$$

$$\hat{\ }_i = 1 + \frac{1}{0} \exp \hat{\ }_0 + \hat{\ }_{1i} + \cdots + \hat{\ }_{Li}$$

$$\sum_{i \in S} {}_i \hat{\ }_i(\boldsymbol{\varphi}) \quad (\boldsymbol{\beta}; {}_i \quad {}_i) = 0$$

$= {}_0\, \mathcal{H} \quad {}_1$

$\hat{\alpha}$

$$(\boldsymbol{\beta}; {}_i \quad {}_i) \quad {}_{i1} \quad {}_i = \sum_{k-1}^{L} \quad {}_l({}_{i1} \quad {}_i)$$

3.2.2 Standard Errors of Estimates

$\boldsymbol{\varphi}$

$\boldsymbol{\beta}$

${}_i = \mathbb{I}(\in \quad {}_0)$

$$\hat{\ }_1(\boldsymbol{\varphi}) = \sum_{i \in S} {}_i \quad {}_i(\boldsymbol{\varphi}) - 1 \quad {}_i$$

$$\hat{\ }_2(\boldsymbol{\varphi} \quad) = \sum_{i \in S} {}_i \hat{\ }_i(\boldsymbol{\varphi}) \quad (\boldsymbol{\beta}; {}_i \quad {}_i)$$

${}_i = (1 \quad {}_{1i} \quad \cdots \quad {}_{Li})$

$$_i(\boldsymbol{\varphi}) = 1 + \frac{1}{0} \exp \quad {}_0 + \quad {}_{1i} + \cdots + \quad {}_{Li}$$

$\hat{\boldsymbol{\beta}}$

$$\hat{\ }_1(\boldsymbol{\varphi}) = 0 \quad \hat{\ }_2(\boldsymbol{\varphi} \quad \boldsymbol{\beta}) = 0$$

$\theta = (\varphi \quad \beta)$

$$\hat{\ }(\theta) = \begin{pmatrix} \hat{\ }_1(\varphi) \\ \hat{\ }_2(\varphi \quad \beta) \end{pmatrix}$$

$\hat{\theta}$

$$(\hat{\theta}) = \ ^{-1} \quad (\hat{\ }) \ ^{-1\prime} \qquad = \quad \frac{}{\theta} \hat{\ }(\theta)$$

$(\hat{\theta})$

$$\tilde{\ } = \frac{}{} \hat{\ }(\ )\Big|_{\theta-\hat{\theta}} \qquad \tilde{\ }(\hat{\ }) = \sum_{i \ S} (\tilde{\ }_i - \bar{\tilde{\ }}_i)(\tilde{\ }_i - \bar{\tilde{\ }}_i)$$

$(\hat{\ })$ $\hat{\ } = (\hat{\ } \quad \hat{\beta})$

$$\tilde{\ }_i = \begin{pmatrix} i \quad i(\hat{\ }) - 1 \quad i \\ i \hat{\ } i(\hat{\ }) \quad (\hat{\beta}; \ i \quad i) \end{pmatrix} \qquad \bar{\tilde{\ }}_i = \frac{1}{} \sum_{i \ S} \tilde{\ }_i$$

3.3 Estimation scheme

$\mathcal{H} = \quad 1(\ i1 \quad i) \ \cdots \ L(\ i1$

Chapter 4

Simulation study

1

- x_{Ai}
- x_{Gi}
- x_{Ti}

Step 3: [illegible]

[illegible] β_1 [illegible] β_2 [illegible]

- Naive model [illegible]
- Traditional model [illegible] $\mathcal{H}$ [illegible]
- Full model [illegible]
- Boosting model [illegible] β_1 [illegible] $\hat{\eta}_i = \exp(\cdot_{i1}\hat{\beta}_1)$ [illegible] $\log \eta_i$ [illegible] $\hat{\beta}_2$ [illegible]
- Proposed model [illegible]

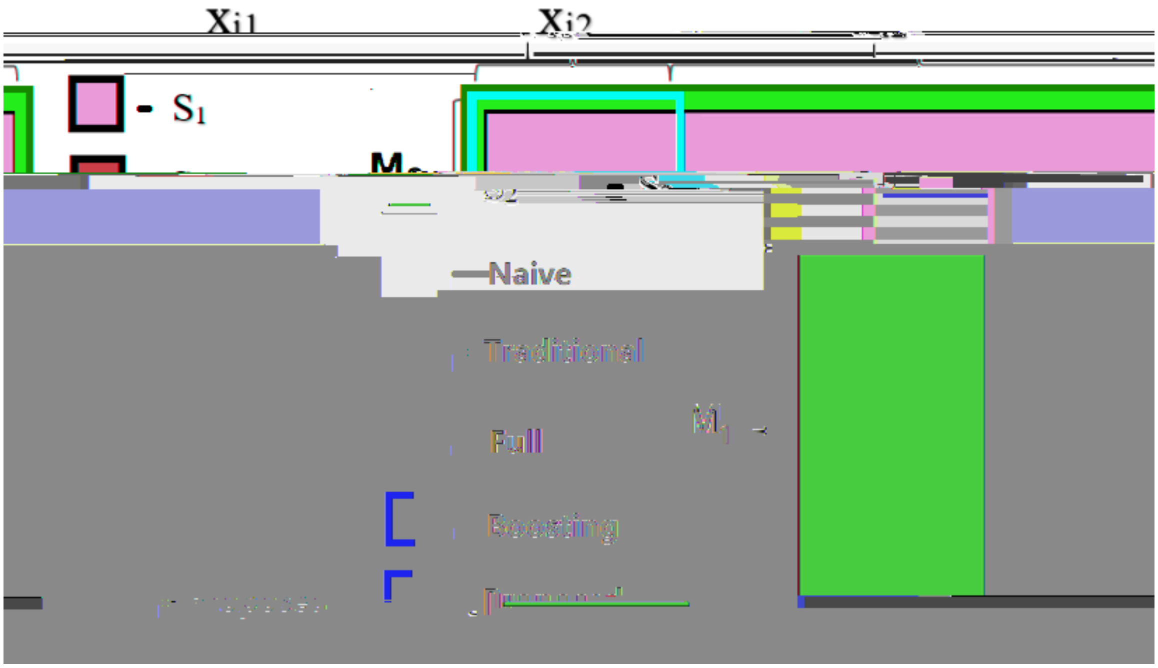
[illegible] th [illegible]

4.3 Evaluation Procedure

[illegible] $= 1\ 000$ [illegible] β_j [illegible]

$$\cdot_j = \frac{1}{\cdot} \sum_{r-1} (\cdot_j - \hat{\cdot}_j^{(r)})$$

$$\cdot_j = \sqrt{\frac{1}{\cdot} \sum_{r-1} (\cdot_j - \hat{\cdot}_j^{(r)})^2}$$



$$\hat{\ }_j = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}_{\{|\beta_j - \hat{\beta}_j^{(r)}| < 1. \cdot \mathrm{SE}(\hat{\beta})_j^{(r)}\}}$$

[illegible]	[illegible]	[illegible]
[illegible]	$\sum_{i \in S_0} [illegible](\boldsymbol{\beta}; [illegible]_i [illegible]_i) = 0$	$[illegible]_i = \exp([illegible]_{i1}\hat{\boldsymbol{\beta}}_1 + [illegible]_{i2}\hat{\boldsymbol{\beta}}_2)$
[illegible]	$\sum_{i \in S} [illegible](\boldsymbol{\beta}_1; [illegible]_{i1} [illegible]_i) = 0$	$[illegible]_i = \exp([illegible]_{i1}\hat{\boldsymbol{\beta}}_1)$
[illegible]	$\sum_{i \in S^*} [illegible](\boldsymbol{\beta}; [illegible]_i [illegible]_i) = 0$	$[illegible]_i = \exp([illegible]_{i1}\hat{\boldsymbol{\beta}}_1 + [illegible]_{i2}\hat{\boldsymbol{\beta}}_2)$
[illegible]	$\sum_{i \in S} [illegible](\boldsymbol{\beta}_1; [illegible]_{i1} [illegible]_i) = 0$	$[illegible]_i = \hat{\eta}_i \exp([illegible]_{i2}\hat{\boldsymbol{\beta}}_2)$
	$\sum_{i-1}^{M_0} [illegible]_i - \hat{\eta}_i \exp([illegible]_{i2}\boldsymbol{\beta}_2) [illegible]_{i2} = 0$	[illegible] $\hat{\eta}_i = \exp([illegible]_{i1}\hat{\boldsymbol{\beta}}_1)$
	[illegible] $\hat{\eta}_i = \exp([illegible]_{i1}\hat{\boldsymbol{\beta}}_1)$ [illegible] $_0$	[illegible]
[illegible]	$\sum_{i \in S} [illegible]_i \hat{[illegible]}_i(\boldsymbol{\varphi}) [illegible](\boldsymbol{\beta}; [illegible]_i [illegible]_i) = 0$	$[illegible]_i = \exp([illegible]_{i1}\hat{\boldsymbol{\beta}}_1 + [illegible]_{i2}\hat{\boldsymbol{\beta}}_2)$
	[illegible] $\hat{[illegible]}_i(\boldsymbol{\varphi})$ [illegible]	
	[illegible] $_i = \mathbb{I}([illegible] \in [illegible]_0)$	

[illegible]

4.4 Results

[illegible] N T B F [illegible] P [illegible] $_0$ [illegible] $\hat{[illegible]}_0$ [illegible] $\hat{[illegible]}_T$ [illegible] $_0$ [illegible] $_0$ [illegible]

[illegible] $\beta_2 = (\beta_T)$ [illegible] $\beta_1 = (\beta_0 \ \beta_{A1} \ \beta_{A2} \ \beta_G)$ [illegible] β_2 [illegible]

	[illegible]					[illegible]					[illegible]				
	N	T	B	F	P	N	T	B	F	P	N	T	B	F	P
Random selection															
β_0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_{A1}	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_{A2}	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_G	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_T	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
Age selection															
β_0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_{A1}	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_{A2}	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_G	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_T	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
Adverse selection															
β_0		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_{A1}	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_{A2}	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_G	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
β_T	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Random selection			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
Age selection			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
Adverse selection			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

[illegible] Right.turns [illegible]

[illegible]	[illegible]	[illegible]
[illegible]	Duration	[illegible]
	Insured.age	[illegible]
	Insured.sex	[illegible]
	Car.age	[illegible]
	Marital	[illegible]
	Car.use	[illegible]
	Credit.score	[illegible]
	Region	[illegible]
	Annual.miles.drive	[illegible]
	Years.noclaims	[illegible]
	TerritoryEmb	[illegible]
[illegible]	Annual.pct.driven	[illegible]
	Total.miles.driven	[illegible]
	Pct.drive.xxx	[illegible]
	Pct.drive.rush.am	[illegible]
	Pct.drive.rush.pm	[illegible]
	Avgdays.week	[illegible]
	Accel.06miles	[illegible]
	Brake.06miles	[illegible]
	Acbr.others	[illegible]
	Left.turns	[illegible]
	Right.turns	[illegible]
[illegible]	NB_Claim	[illegible]

[illegible]

5.2 Estimation and Evaluation

[illegible]

5.2.1 Estimation

- Random selection
- Age selection $1/(1+\exp(0.03\,\texttt{Insured.age}_i))$
- Adverse selection $1/(1+\exp(\texttt{NB\_Claim}_i))$

$= 500$

$\mathcal{H}_B$

5.2.2 Evaluation

r56j ET qv7 7. TD5.5

$$j = \frac{1}{r-1}$$

[illegible]

- [illegible] 0 [illegible] 1 [illegible] 0 [illegible]
- [illegible]
- [illegible]

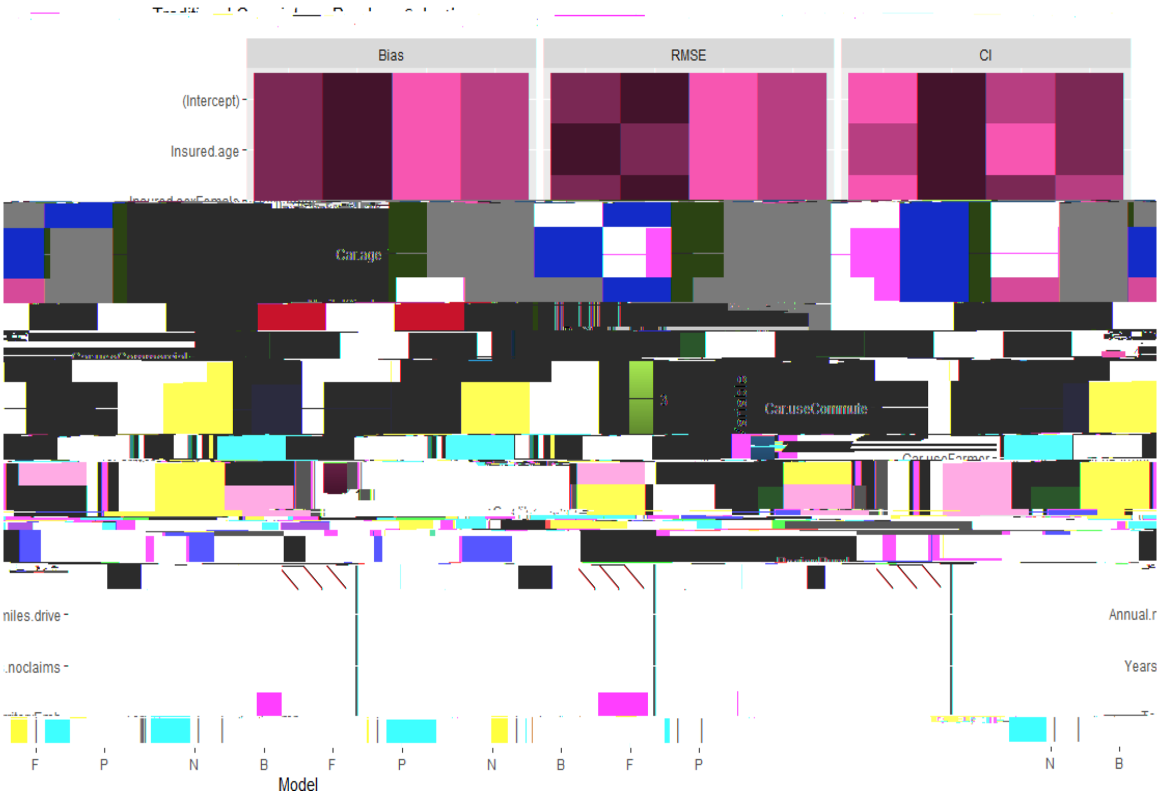
[illegible]

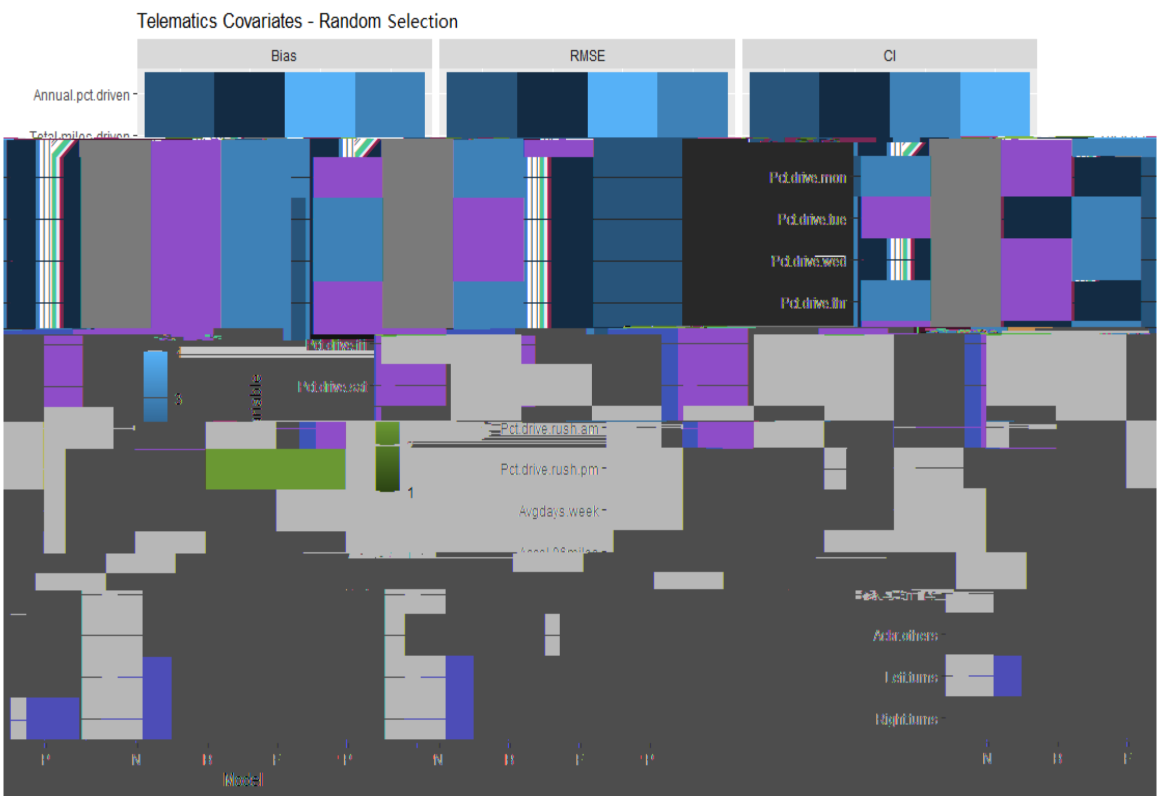
[illegible]

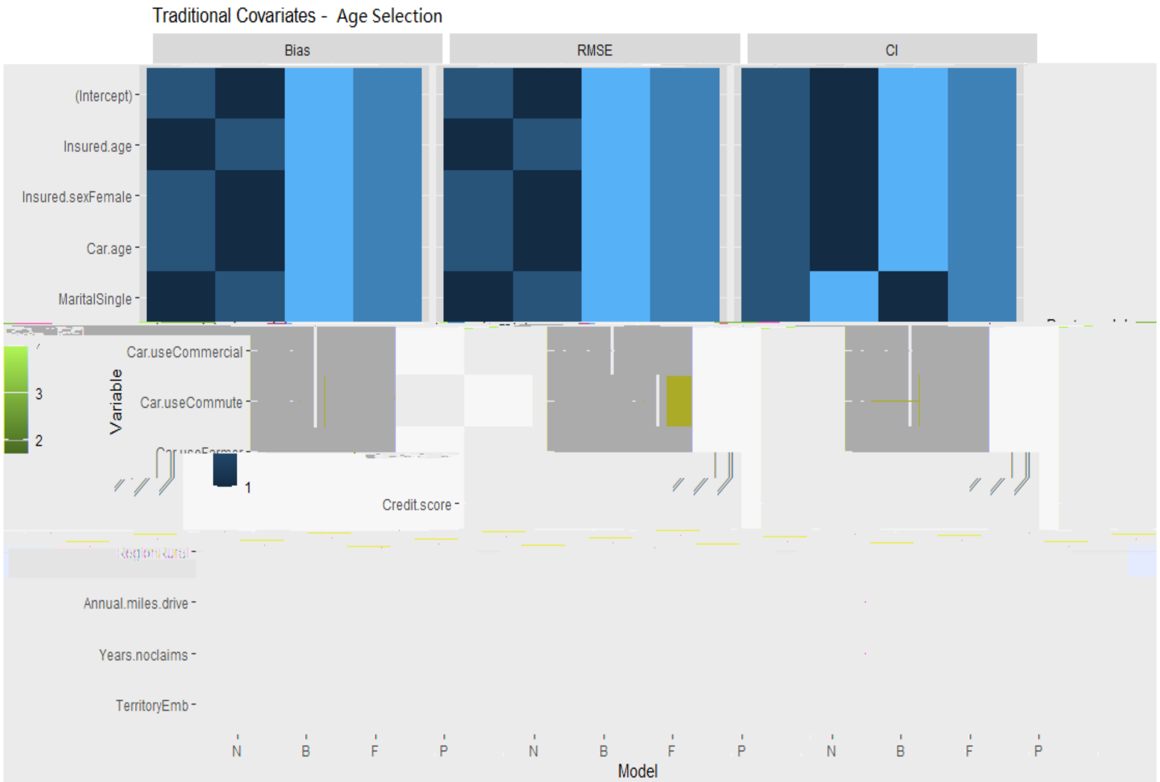
Random selection					
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
Age selection					
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
Adverse selection					
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	

[illegible]

[illegible]


Bias
RMSE
CI
(Intercept)
Insured.age
Car.age
Car.useCommute
Annual.r
Years
F
P
N
B
Model

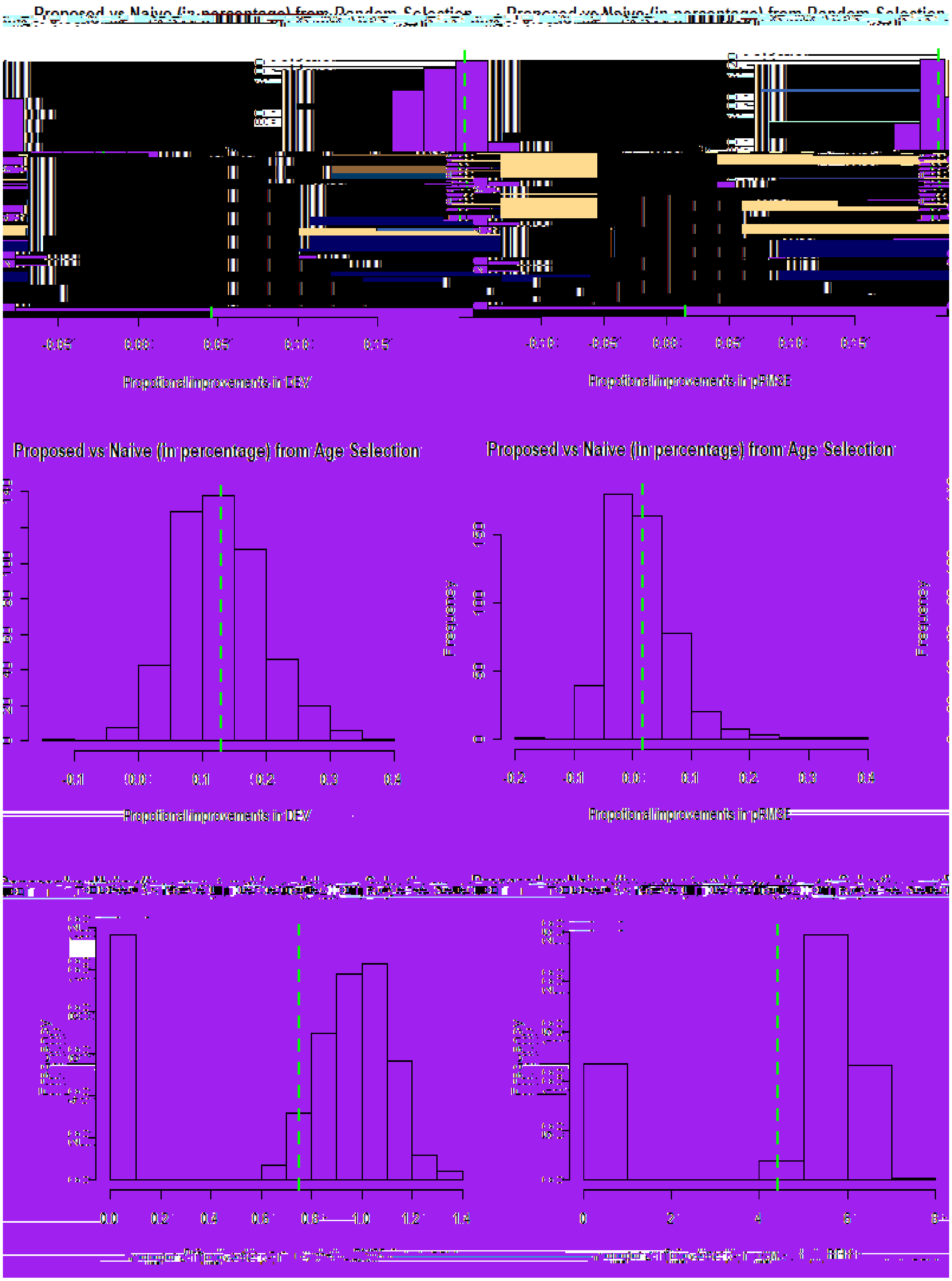

Telematics Covariates - Random Selection
Bias
RMSE
CI
Annual.pct.driven
Pct.drive.mon
Pct.drive.tue
Pct.drive.wed
Pct.drive.thr
Pct.drive.sat
Pct.drive.rush am
Pct.drive.rush.pm
Avgdays.week
Left.turns
Night.time


Traditional Covariates - Age Selection
Bias
RMSE
CI
(Intercept)
Insured.age
Insured.sexFemale
Car.age
MaritalSingle
Car.useCommercial
Car.useCommute
Credit.score
Annual.miles.drive
Years.noclaims
TerritoryEmb
Variable
3
2
1
N
B
F
P
Model


CI
Bias
RMSE
Annual.pct.driven
Pct.drive.fri
Best_model
4
Model


Traditional Covariates - Adverse Selection
Bias
RMSE
CI
(Intercept)
Insured.age
Credit.score
Best_model


Telematics Covariates - Adverse Selection
CI
Bias
RMSE
Annual.pct.driven
Total.miles.driven
Pct.drive.mon
Pct.drive.tue
Pct.drive.thr
Pct.drive.fri
Pct.drive.sat
Pct.drive.rush.am
Pct.drive.rush.pm
Avgdays.week
Accel.06miles
Brake.06miles
Acbr.others
Left.turns
Right.turns
Best_model
4
3
2
1
Variable
N
B
F
P
Model


Proposed vs Naive (in percentage) from Random Selection
Proposed vs Naive (in percentage) from Random Selection
Proportional improvements in DEV
Proportional improvements in pRMSE
Proposed vs Naive (in percentage) from Age Selection
Proposed vs Naive (in percentage) from Age Selection
Frequency
Proportional improvements in DEV
Proportional improvements in pRMSE

Bibliography

Categorical data analysis

Accident Analysis and Prevention

Risks

Transportation

Decision Support Systems

Insurance: Mathematics and Economics

Accident Analysis & Prevention

ASTIN Bulletin: The Journal of the IAA

Risk analysis

Journal of Risk and Insurance

Journal of the Royal Statistical Society: Series D (The Statistician)

Scandinavian Actuarial Journal

Procedia Engineering

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

Proceedings of the 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications

Journal of Risk Management

Available at SSRN 3251623

Variance

Insurance: Mathematics and Economics

Casualty Actuarial Society E-Forum, Winter 2011 Volume 2

Canadian Electronic Library https://policycommons.net/artifacts/1189025/distance-based-vehicle-insurance/1742147/

Transportation Research Part A: Policy and Practice

Big Data & Society

Sustainability

Risks

Biometrika

International Journal of Statistics and Systems

Journal of the operational research society

Risks

Accident Analysis and Prevention

Principles and Applications of Narrowband Internet of Things (NBIoT)

Journal of the Royal Statistical Society: Series C (Applied Statistics)

Big Data for Insurance Companies

arXiv e-prints

Appendix A

Results

	N	B	F	P	N	B	F	P	N	B	F	P
Adverse selection												
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Appendix B

Code

```
x4        <- rnorm(I)
lambda    <- exp(-1.3-4*x1 + 3.4*x2 + 0.1*x3 + 0.5*x4)
NB_Claim  <- rpois(I, lambda)
Duration  <- rep(1, I)
fdata     <- as.data.frame(cbind(x1, x2, x3, x4, Duration, NB_Claim))

#for testing
set.seed(j+1000)
test_ind <- sample(1:nrow(fdata), 10000)
forr.data<-fdata[ test_ind,]
trtt.data<-fdata[-test_ind,]
#sampling-  when using a specific sampling method, comment other two sampling sections.
##################################################RS
set.seed(j+2000)
tele_ind <- sample(1:90000, nrow(fdata)*0.1)
ntr <- length(tele_ind)
##################################################NIS(advsel)
#set.seed(j+2000)
#dz <- 1/(1+exp(2*trtt.data$NB_Claim))
#dz <- dz/mean(dz)/9
#dzz <- rbinom(90000, 1, dz)
#tele_ind <- (1:90000)*(dzz==1)
#rm(dz, dzz)
#tele_ind <- tele_ind[tele_ind!=0]
#ntr <- length(tele_ind)
##################################################MAR(agesel)
#set.seed(j+2000)
#dz <- 1/(1+exp(3*trtt.data$x1))
#dz <- dz/mean(dz)/9
#dzz <- rbinom(90000, 1, dz)
#tele_ind <- (1:90000)*(dzz==1)
#rm(dz, dzz)
#tele_ind <- tele_ind[tele_ind!=0]
#ntr <- length(tele_ind)
###################################################
S0  <- trtt.data[ tele_ind,           ]
# A small dataset that contains both telematics and traditional features
S1  <- trtt.data[-tele_ind,-tele_ind,-tele_ind,)
       T*Td(<-)Tj0g14.1220Td[(trtt.data[)-525.004(tele_ind,)]TJ150220Td[(trt.12214_in9D4]TJ5atetelematics)-
```

```r
b_S0 <- as.matrix(cbind(S0[,c(5,1:3)], S0[,6]*S0[,c(5,1:3)]))
b_S  <- as.matrix(cbind(S[ ,c(5,1:3)], S[ ,6]*S[ ,c(5,1:3)]))

#function for optimize using nleqslv()
cal_eqn <- function(parm) {
  result <- colSums(as.vector(1+nrow(S1)/nrow(S0)*exp(parm%*%t(b_S0)))*b_S0)-colSums(b_S)
  return(result)  }

#find for parameters of basis functions
fit2 <- nleqslv(rep(0,8), cal_eqn)

#calculate weights from information projection
w3   <- 1+nrow(S1)/nrow(S0)*exp(b_S0 %*% fit2$x)
#############################
#combine weights to S0
SS6<-cbind(S0,w3)

#fitted the model with ws
glm.freq.S3 <- glm(NB_Claim ~ .-Duration-w3,offset=log(Duration),
                   weights= w3,  data=SS6, family=poisson())
x_S0 <- model.matrix(glm.freq.S3)

#coef of proposed model
prop2_coef[j,] <- summary(glm.freq.S3)$coefficients[,1]

# sandwich formula for variance estimation
Ui <- cbind(c(as.vector(w3)-1, rep(-1, nrow(S1)))* b_S,
  c(w3*(SS6$NB_Claim-fitted(glm.freq.S3)), rep(0, nrow(S1)))*as.matrix(S[,c(5,1:4)]))

Ui <- Ui - rep(colMeans(Ui), each=nrow(S))
V_U  <- (  t(Ui) %*% Ui)
tau  <- rbind(cbind(t(b_S0) %*% (as.vector(w3-1)*b_S0),
        matrix(0, ncol=ncol(x_S0), nrow=ncol(b_S0))),
  cbind(t(x_S0) %*% (as.vector(w3-1)*(SS6$NB_Claim-fitted(glm.freq.S3))*b_S0),
        -t(x_S0) %*% (as.vector(w3*fitted(glm.freq.S3))*x_S0) ))
invtau <- solve(tau)
prop2_stde[j,]  <- sqrt(diag(invtau %*% V_U %*% invtau))[-(1:ncol(b_S0))]
##########################boosting model
glm.freq.boost  <- glm(NB_Claim ~ x4-1,data=S0,
            offset=log(Duration)+predict(glm.freq.trad, S0), family=poisson())
#coefficients and SE
boost_coef[j,] <- c(summary(glm.freq.trad )$coefficients[,1],
                    summary(glm.freq.boost)$coefficients[,1])
boost_stde[j,] <- c(summary(glm.freq.trad )$coefficients[,2],
                    summary(glm.freq.boost)$coefficients[,2])
################################try forecast
pred.naive <- predict(glm.freq.naive, newdata = forr.data, type="response")
pred.full  <- predict(glm.freq.full , newdata = forr.data, type="response")
pred.S3    <- exp(as.matrix(forr.data[,c(5,1:4)]) %*% prop2_coef[j,])
pred.trad  <- predict(glm.freq.trad , newdata = forr.data, type="response")
pred.boost <- pred.trad * exp(coef(glm.freq.boost)*forr.data$x4)

#remove datasets for this split
rm(tele_ind, test_ind)

#RMSE
RMSEs[j,-4] <- sqrt(c(
  mean((forr.data$NB_Claim-pred.naive)^2),
```

```r
      mean((forr.data$NB_Claim-pred.trad )^2),
      mean((forr.data$NB_Claim-pred.full )^2),
      mean((forr.data$NB_Claim-pred.S3   )^2),
      mean((forr.data$NB_Claim-pred.boost)^2)))

    #MAE
    MAEs[j,-4] <- c(
      mean(abs(forr.data$NB_Claim-pred.naive)),
      mean(abs(forr.data$NB_Claim-pred.trad )),
      mean(abs(forr.data$NB_Claim-pred.full )),
      mean(abs(forr.data$NB_Claim-pred.S3   )),
      mean(abs(forr.data$NB_Claim-pred.boost)))

    #DEV
    DEVs[j,-4] <- c(
      Poisson.Deviance(pred.naive, forr.data$NB_Claim),
      Poisson.Deviance(pred.trad , forr.data$NB_Claim),
      Poisson.Deviance(pred.full , forr.data$NB_Claim),
      Poisson.Deviance(pred.S3   , forr.data$NB_Claim),
      Poisson.Deviance(pred.boost, forr.data$NB_Claim))

  })

#summarizing the outputs
colMeans(RMSEs)
colMeans(MAEs)
colMeans(DEVs)

#true coeffients used for data generation
true_coef      <- c(-1.3, -4, 3.4, 0.1, 0.5)

#bias of each estimator
bias_naive <- true_coef - colMeans(naive_coef)
bias_trad  <- true_coef - colMeans(trad_coef)
bias_prop2 <- true_coef - colMeans(prop2_coef)#proposed
bias_boost <- true_coef - colMeans(boost_coef)
bias_full  <- true_coef - colMeans(full_coef)

#RMSE of estimates
rmse_naive <- sqrt(colMeans((naive_coef-rep(true_coef, each=J))^2))
rmse_trad  <- sqrt(colMeans((trad_coef -rep(true_coef, each=J))^2))
rmse_prop2 <- sqrt(colMeans((prop2_coef-rep(true_coef, each=J))^2))#proposed
rmse_boost <- sqrt(colMeans((boost_coef-rep(true_coef, each=J))^2))
rmse_full  <- sqrt(colMeans((full_coef -rep(true_coef, each=J))^2))

#CI of estimator
naive_90CI <- colMeans((naive_coef-1.645*naive_stde<rep(true_coef, each=J))*
                         (naive_coef+1.645*naive_stde>rep(true_coef, each=J))*1)
trad_90CI <- colMeans((trad_coef -1.645*trad_stde<rep(true_coef, each=J))*
                        (trad_coef +1.645*trad_stde>rep(true_coef, each=J))*1)
prop2_90CI <- colMeans((prop2_coef-1.645*prop2_stde<rep(true_coef, each=J))*
                         (prop2_coef+1.645*prop2_stde>rep(true_coef, each=J))*1,
                       na.rm=TRUE)#proposed
boost_90CI <- colMeans((boost_coef-1.645*boost_stde<rep(true_coef, each=J))*
                         (boost_coef+1.645*boost_stde>rep(true_coef, each=J))*1)
full_90CI <- colMeans((full_coef -1.645*full_stde<rep(true_coef, each=J))*
                        (full_coef +1.645*full_stde>rep(true_coef, each=J))*1)
```

```r
#### Preliminary analysis ####
#naive model
glm.freq.naive <- glm(NB_Claim ~ .-Duration,offset=log(Duration),data=S0, family=poisson())
#full model
glm.freq.full  <- glm(NB_Claim ~ .-Duration,offset=log(Duration),data=S , family=poisson())
#traditional model
glm.freq.trad  <- glm(NB_Claim ~ .-Duration,offset=log(Duration),
                      data=S[, c(1:13, 30)], family=poisson())
```

```
                         data=S0, offset=log(Duration)+predict(glm.freq.trad, S0)
                         , family=poisson())

  #coefficients and SE
  boost_coef[j,] <- c(summary(glm.freq.trad )$coefficients[,1],
                      summary(glm.freq.boost)$coefficients[,1])
  boost_stde[j,] <- c(summary(glm.freq.trad )$coefficients[,2],
                      summary(glm.freq.boost)$coefficients[,2])
  ################################try forecast
  pred.naive <- predict(glm.freq.naive, newdata = forr.data, type="response")
  pred.full  <- predict(glm.freq.full , newdata = forr.data, type="response")
  pred.S3    <- exp(as.matrix(forr.data[,c(2:29)]) %*%
             propd_coef[j,2:29]+propd_coef[j,1]+ log(forr.data[,1]))
  pred.trad  <- predict(glm.freq.trad , newdata = forr.data, type="response")
  pred.boost <- pred.trad * exp(as.matrix(forr.data[14:29])%*%coef(glm.freq.boost))

  #remove datasets for this split
  rm(tele_ind, test_ind)

  #RMSE
  RMSEs[j,] <- sqrt(c(
    mean((forr.data$NB_Claim-pred.naive)^2),
    mean((forr.data$NB_Claim-pred.trad )^2),
    mean((forr.data$NB_Claim-pred.full )^2),
    mean((forr.data$NB_Claim-pred.S3   )^2),
    mean((forr.data$NB_Claim-pred.boost)^2)))

  #DEV
  DEVs[j,] <- c(
    Poisson.Deviance(pred.naive, forr.data$NB_Claim),
    Poisson.Deviance(pred.trad , forr.data$NB_Claim),
    Poisson.Deviance(pred.full , forr.data$NB_Claim),
    Poisson.Deviance(pred.S3   , forr.data$NB_Claim),
    Poisson.Deviance(pred.boost, forr.data$NB_Claim))
})
```

Appendix C

Basic Setup of Proposed Method

$$\perp X_2 \mid X_1$$

$(X_1, X_2) = (X_1)$

$(X_1) = (\,_1(X_1), \ldots, \,_L(X_1))$

$$\perp X_2 \mid X_1$$

(X_1)

$$(X_1) = \mathrm{E}\,(X_1)\,(X_1) = 1$$

$$\frac{1}{0}\sum_{i=1}^{M} {}_i\,(X_{i1})\,(X_{i1}) = \frac{1}{1}\sum_{i=1}^{M}(1 - {}_i)\,(X_{i1})$$

(X_1)

$$\hat{}_{SPS}(\,) = \frac{1}{}\sum_{i=1}^{M} {}_i\left[1 + \frac{1}{0}\,(X_{i1})\right](\,; X_{i1}, X_{i2})$$

(X_1) (X_1)

$$\mathrm{E}\,(X_1)\,(X_1)\,X_{i2} = 1 = \mathrm{E}\,(X_1)\,(X_1) = 1$$

E