



UNIVERSITAT DE  
BARCELONA

# Motor insurance with telematics driving data

Montserrat Guillen

Seminar organized by ETH Zurich, RiskLab

September 16, 2022





Ana M. Pérez-Marín



Jens P. Nielsen

Ainoa Murillo-López



Giselle Aguer



Leandro Masello & co-authors



# Contents

1. Introduction
2. Methods
3. Case Studies
4. Conclusions & take-home

# Contents

**1. Introduction**

2. Methods

3. Case Studies

4. Conclusions & take-home

# What is not telematics car driving data?

- Classical covariates:
  - Car-related features  
Type of car, brand, vehicle model, horsepower, etc.
  - Driver related features  
Age, gender, health condition, children, occupation, etc.
  - Insurance contract information  
Type of contract, duration and other features
  - Annual mileage, vehicle use, claims experience, etc.
- In general, 50 potential covariates are typically used in classical motor insurance pricing



# Super Easy Octo Telematics

# How do raw telematics data look like?

```
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:38,"2017-06-08",10.23,0,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:39,"2017-06-08",9.45,9.45,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:40,"2017-06-08",8.83,0.053333,253,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:41,"2017-06-08",8.41,-0.606667,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:42,"2017-06-08",8.29,-0.386667,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:43,"2017-06-08",8.72,-0.036665,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:44,"2017-06-08",8.75,0.113333,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:45,"2017-06-08",8.42,0.043333,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:46,"2017-06-08",7.95,-0.256667,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:47,"2017-06-08",7.85,-0.3,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:48,"2017-06-08",7.92,-0.166667,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:49,"2017-06-08",8.5,0.183334,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:50,"2017-06-08",9.17,0.44,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:51,"2017-06-08",10.13,0.736667,254,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:52,"2017-06-08",10.76,0.753333,253,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:53,"2017-06-08",11.14,0.656667,253,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:54,"2017-06-08",11.44,0.436667,253,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:55,"2017-06-08",11.64,0.293333,253,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:56,"2017-06-08",11.22,0.026667,253,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:57,"2017-06-08",11.11,-0.11,253,19
-4c81-11e7-bd41-0a7942277526",2017-06-08 19:21:58,"2017-06-08",11.21,-0.143333,253,19
```

Vehicle ID, Timestamp, Date, Distance, Acceleration, Road type, County

Telematics raw data file. Sun et al. (2021)

# What is, then, telematics data?

- Global Positioning Signal (GPS) –not always-
- Speed, acceleration, braking, and turn intensity
- Vehicle sensors and cameras
- Engine information
- Timestamp and mileage
- Traffic rules and context conditions
- Passengers, distractions, smartphone use
- High-frequency time series information recorded during driving
- A challenge? The volume of raw data. What are the relevant summaries? How much monitoring is enough?



# Questions

- Insurance companies collect **telematics data** about drivers' **exposure to traffic** (distance driven, usage frequency and type of road) and their **driving behavior** (excess speed, aggressiveness, operating hours). In addition, **context information** (traffic conditions, weather) can also be accessed.
  - This information can be used to:
    - **improve the insurance ratemaking process.**
    - **promote safe driving.**
- (1) How are pay-per-mile **insurance schemes** be designed?
  - (2) How can near-miss (**risky event**) telematics be used to **identify risky drivers**?
  - (3) Does risk analytics and percentile charts help **monitoring drivers**?

# What has been **written so far** about telematics car driving data?

- Transportation Literature
  - Vehicle emissions, energy consumption and traffic impact.
  - Driving behavior and accidents.
- Insurance Literature (**Usage Based Insurance UBI**)
  - The beginnings: **PAYD**, mileage and accidents
  - Driving habits, skills and behavior:  
Pay-as-you-drive → pay-how-you-drive
  - The problem of low frequency of claims:  
A new concept: **near-miss incidents**

# Actuarial literature & telematics driving data

- Telematics ratemaking **recent research**:

Barry & Charpentier (2020) -personalization/pooling-,  
Geyer, Kremslehner & Mürmann (2020)–contract choice-  
Eling & Kraft (2020) – 52 articles in 20 years-,  
So, Boucher & Valdez(2021) – synthetic data set -,  
Duval, Boucher & Pigeon(2021) -3 months of telematics data is enough-  
...and lately a lot on Machine Learning.  
Gao, Wang & Wüthrich (2022) – data sources interact-  
Richman & Wüthrich (2022) – improves interpretation-  
Fung, Tzougas & Wüthrich (2022) – claim severity-

- Key **methodological questions**:

- Time frame (yearly, monthly, weekly rates)
- Distance driven (linear or log-linear)
- Driving style (which indicators? which conditions?)
  - Urban/Non urban; Younger drivers/Older drivers; Type of vehicle
- Score/Classify drivers (**Wüthrich, Gao & Wang**)

- The quality of **telematics data**:

– Raw data are not always as good as they should be

(**sensor errors, clock errors, inertial measurement failures, summertime/wintertime issues, GPS blanks,...**)<sup>11</sup>


# Telematics data: today in 2022



[MORE DETAILS](#)

# Contents

1. Introduction
- 2. Methods**
3. Case Studies
4. Conclusions & take-home

An aerial photograph of the Tesla Gigafactory Texas. The building is a large, rectangular structure with a flat roof covered in solar panels. The solar panels are arranged in a grid pattern, with several large, white, stylized 'T' logos interspersed. The building is surrounded by a parking lot with many cars, and there are trees and other buildings in the background. The text 'GIGAFACTORY TEXAS' is overlaid in the upper right corner.

GIGAFACTORY  
TEXAS

Tesla's Safety Score

# Tesla's Five Factors



Forward Collision  
Warnings per  
1,000 Miles\*

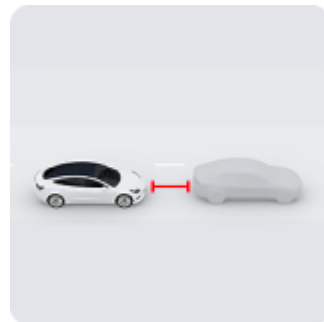
\*capped 102



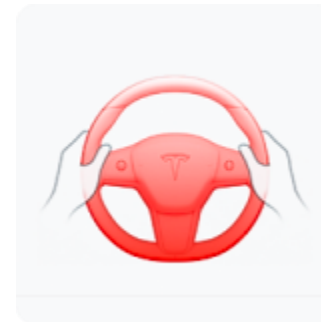
Hard Braking  
[>3 m/s<sup>2</sup>, Prop  
0.3G/0.1G]



Aggressive Turning  
[Prop. 0.4G/0.2G  
lateral acceleration]



Unsafe Following  
[Prop. 1sec/3sec  
Speed >50mph (80Km/h)]



Forced Autopilot Disengagement  
[After 3 warnings of inattentive,  
no hands on the wheel]



# <https://www.tesla.com/support/safety-score>

## **Forward Collision Warnings per 1,000 Miles**

Forward Collision Warnings are audible and visual alerts provided to you, the driver, in events where a possible collision due to an object in front of the vehicle is considered likely without your intervention. Events are captured based on the 'medium' Forward Collision Warning sensitivity setting regardless of your user's setting in the vehicle. Forward Collision Warnings are incorporated into the Safety Score formula at a rate per 1,000 miles. The value is capped at 101.9 per 1,000 miles in the Safety Score formula.

## **Hard Braking**

Hard braking is defined as backward acceleration, measured by your Tesla vehicle, in excess of 0.3g. This is the same as a decrease in the vehicle's speed larger than 6.7 mph, in one second. Hard braking is introduced into the Safety Score formula as the proportion of time (expressed as a percentage) where the vehicle experiences backward acceleration greater than 0.3g relative to the proportion of time where the vehicle experiences backward acceleration greater than 0.1g (2.2 mph in one second). Hard braking while on Autopilot is not factored into the Safety Score formula. The percentage shown in the app is the percentage of manual braking that is done with excessive force when driving and Autopilot is not engaged. The value is capped at 7.4% in the Safety Score formula.

## **Aggressive Turning**

Aggressive turning is defined as left/right acceleration, measured by your Tesla vehicle, in excess of 0.4g. This is the same as an increase in the vehicle's speed to the left/right larger than 8.9 mph, in one second. Aggressive turning is introduced into the Safety Score formula as the proportion of time (expressed as a percentage) where the vehicle experiences lateral acceleration greater than 0.4g, in either the left or right direction, relative to the proportion of time where the vehicle experiences acceleration greater than 0.2g (4.5 mph in one second), in either the left or right direction. Aggressive turning while on Autopilot is not factored into the Safety Score formula. The percentage shown in the app is the percentage of turning that is done with excessive force when driving and Autopilot is not engaged. The value is capped at 17.1% in the Safety Score formula.

## **Unsafe Following**

Your Tesla vehicle measures its own speed, the speed of the vehicle in front and the distance between the two vehicles. Based on these measurements, your vehicle calculates the number of seconds you would have to react and stop if the vehicle in front of you came to a sudden stop. This measurement is called "headway." Unsafe following is the proportion of time where your vehicle's headway is less than 1.0 seconds relative to the time that your vehicle's headway is less than 3.0 seconds. Unsafe following is only measured when your vehicle is traveling at least 50 mph and is incorporated into the Safety Score formula as a percentage. Unsafe following while on Autopilot is not factored into the Safety Score formula. The percentage shown in the app is the percentage of unsafe following when driving and Autopilot is not engaged. The value is capped at 60.0% in the Safety Score formula.

## **Forced Autopilot Disengagement**

The Autopilot system disengages for the remainder of a trip after you, the driver, have received three audio and visual warnings. These warnings occur when your Tesla vehicle has determined that you have removed your hands from the steering wheel and have become inattentive. Forced Autopilot Disengagement is introduced into the Safety Score formula as a 1 or 0 indicator. The value is 1 if the Autopilot system is forcibly disengaged during a trip, and 0 otherwise.



# Tesla's Safety Score

Predicted Collision Frequency (PCF) =

0.68 x

1.01 Forward Collision Warnings per 1,000 Miles x

1.13 Hard Braking x

1.02 Aggressive Turning x

1.00 Unsafe Following Time x

1.32 Autopilot Disengagement

The current formula was derived based on statistical modeling using **6 billion miles of fleet data**. Tesla expects to make changes to the formula in the future as more customer and data insights are gained

The PCF is converted into a 0 to 100 Safety Score using the following formula:

**Safety Score** =  $115.382324 - 22.526504 \times \text{PCF}$

# Tesla's Safety Score in log link

**Predicted Collision Frequency (PCF)** =  $\exp\{-0,166 +$   
 $0,006$  Forward Collision Warnings per 1,000 Miles +  
 $0,052$  Hard Braking +  
 $0,008$  Aggressive Turning +  
 $0,001$  Unsafe Following Time  
 $0,120$  Autopilot Disengagement}

Safety Score <sup>Beta</sup>

Based on driving behavior for  
Oct 1, 2021 - Oct 30, 2021



PCF=1,13

Safety Score =  
 $115.38 - 22.53 \times \text{PCF}$

# Yearly Accident frequency to Safety Score

PCF / year	Safety Score
0.03	109
0.06	102
0.07	100
0.08	97
0.09	95
0.10	93
0.12	88
0.14	84
0.20	70

# Is Tesla's Safety Score complete?

- No information on **driver's characteristics**
- No information on **vehicle**
- No information on **external factors**
  - **Weather**
  - **Traffic congestion**
  - **Road type**
  - **Time of day / weekday or weekend**
  - **--- Performance relative to other drivers.**

# An overview of methods

Claims

- **Count data regression** model
- Count data regression panel
- GLM / GAM
- Machine learning approaches

Near misses

- Correlate with claims & reveal information
- New instruments to score drivers

Prevention

- Predictive models of **accident risk**
- Risk maps, driving pulse diagrams (DPD), **percentile charts**

# Notation and classical Poisson model specification (timeframe: yearly data)

- $Y_i$  number of claims at fault policy  $i$ ,  $i = 1, \dots, n$
- $T_i$  risk exposure, offset for policy  $i$
- $x_i, z_i$  vectors of ratemaking factors (traditional  $x_i$ , telematics  $z_i$ )
- A common assumption then is that the numbers of claims  $Y_i$  are independent across all policy holders and they can be modeled by a Poisson regression model

$$\begin{aligned} E(Y_i | x_i, z_i, T_i) &= T_i \exp(x_i' \beta + z_i' \alpha) = \\ &= T_i \exp(x_i' \beta) \exp(z_i' \alpha) = \\ &= \mu(x_i, z_i, T_i) \end{aligned}$$

# Poisson deviance loss

$$L(\hat{\mu}(x_i, z_i, T_i), \mathcal{T}) =$$

$$\frac{2}{|\mathcal{T}|} \sum_{\substack{i \in \mathcal{T} \\ Y_i \neq 0}} Y_i \left( \frac{\hat{\mu}(x_i, z_i, T_i)}{Y_i} - 1 - \log\left(\frac{\hat{\mu}(x_i, z_i, T_i)}{Y_i}\right) \right) +$$

$$\frac{2}{|\mathcal{T}|} \sum_{\substack{i \in \mathcal{T} \\ Y_i = 0}} 2 \cdot \hat{\mu}(x_i, z_i, T_i)$$

$\mathcal{T}$  is the test data set

# Model Boosting: formulas

$$E(Y_i | x_i, z_i, T_i) = T_i \exp(x_i' \beta) \rho(z_i) = \mu(x_i, z_i, T_i)$$

- Two-step approach of **first fitting a GLM** and then **building the telematics risk factor around this GLM** corresponds to the combined actuarial neural network (CANN) model proposed by Wüthrich and Merz (2019).
- Gao et al. (2022) interpret it by studying the network weights and find that **hard braking in low speeds contributes most to a high telematics risk factor**.



# Model Boosting: formulas, with more telematics information

$$E(Y_i | x_i, z_i, u_i, T_i) = T_i \exp(x_i' \beta) \rho(z_i) \varphi(u_i)$$

- With estimated  $\hat{\beta}$  and  $\hat{\rho}(\cdot)$ , then the second telematics risk factor  $\varphi(\cdot)$  is modelled.

# Telematics data by trip data

- Take some **risky drivers** and some **safe drivers**.
- Take the series of trip data for these drivers.
- Construct a **classifier** from these trips.

---

- **Classify all trips** by all drivers based on telematics data:

$$\hat{\psi}(z_{i,j}), i = 1, \dots, n; j = 1, \dots, J_i$$

- Define a **score for each driver**:

$$\bar{\psi}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} \hat{\psi}(z_{i,j})$$

# Telematics trip score in the Poisson model specification

Starting from the classical approach:

$$E(Y_i | x_i, z_i, T_i) = T_i \exp(x_i' \beta + z_i' \alpha) = \\ T_i \exp(x_i' \beta) \exp(z_i' \alpha)$$

Insert the driver's score based on trips or a smoothed credibility version:

$$E(Y_i | x_i, z_i, T_i) = \\ T_i \exp(x_i' \beta) \exp(\alpha_0 + \alpha_1 \bar{\psi}_i)$$

Gao, Meng, Wüthrich (2022) find poorer out-of-sample prediction compared to the v-a heatmap

# Panel binary model specification (timeframe: weekly data)

- $Y_{it}$  binary (claim at fault) policy  $i$ , week  $t$ ,  
 $i = 1, \dots, n \quad t = 1, \dots, W_i$
- $T_{it}$  risk exposure offset for policy  $i$ , week  $t$ ,  
(days?)
- $x_i, z_{it}$  vectors of ratemaking factors (traditional  $x_i$   
, telematics  $z_{it}$ )
- We assume a panel structure where  $Y_{it}$  are  
independent across all policy holders. If there is  
independence over time:

$$\begin{aligned} E(Y_{it} | x_i, z_{it}, T_{it}) &= \mu(x_i, z_{it}, T_{it}) \\ &= \text{Prob}(Y_{it} = 1 | x_i, z_{it}, T_{it}) = p_{it} \end{aligned}$$

# Panel binary model specification (timeframe: weekly data)

- Consider all information to (t-1),  $\mathbf{E}_{t-1}$ :

$$Prob(Y_{it} = 1 | x_i, z_{it}, T_{it}, \mathbf{E}_{t-1}) = p_{it}$$

- We assume a panel structure where  $Y_{it}$  are independent across all policy holders, but they have an autoregressive behavior within the same policy holder.

$$p_{it} = \kappa(p_{i(t-1)} - \theta_i - \xi_{(t-1)}) + \eta_{it} + \theta_i + \xi_t$$

# Contents

1. Introduction
2. Methods
- 3. Case Studies**
4. Conclusions & take-home

# CASE STUDY I

- Pricing with near-misses
- Contextual data



NEAR-MISSES



# What is a *near-miss*?

Near-crash, risky event

- A **near-miss** is a term borrowed from aviation safety – a situation in which **an accident is narrowly avoided**, such as when a driver brakes suddenly in order to avoid a crash (Arai et al., 2001).

**Near-misses** (or incidents) have been shown to be **correlated** with claims in auto insurance

Ma, Y. L., Zhu, X., Hu, X. and Chiu, Y. C. (2018). *The use of context-sensitive insurance telematics data in auto insurance ratemaking*, **Transportation Research Part A** 113, 243–258.

Guillen, M. et al. (2021) Near-miss telematics in motor insurance. **Journal of Risk and Insurance** (OPEN ACCESS)

<https://onlinelibrary.wiley.com/doi/epdf/10.1111/jori.12340>

# Examples: near-misses

- **Acceleration:**  $>6\text{m/s}^2$ , (Hynes & Dickey, 2008).
- **Braking:**  $<-6\text{m/s}^2$
- **Dangerous Turns:** speed combined with angle
- **Use of smart phone while driving**

North American Actuarial Journal (2019) we proposed modeling *near-miss events*

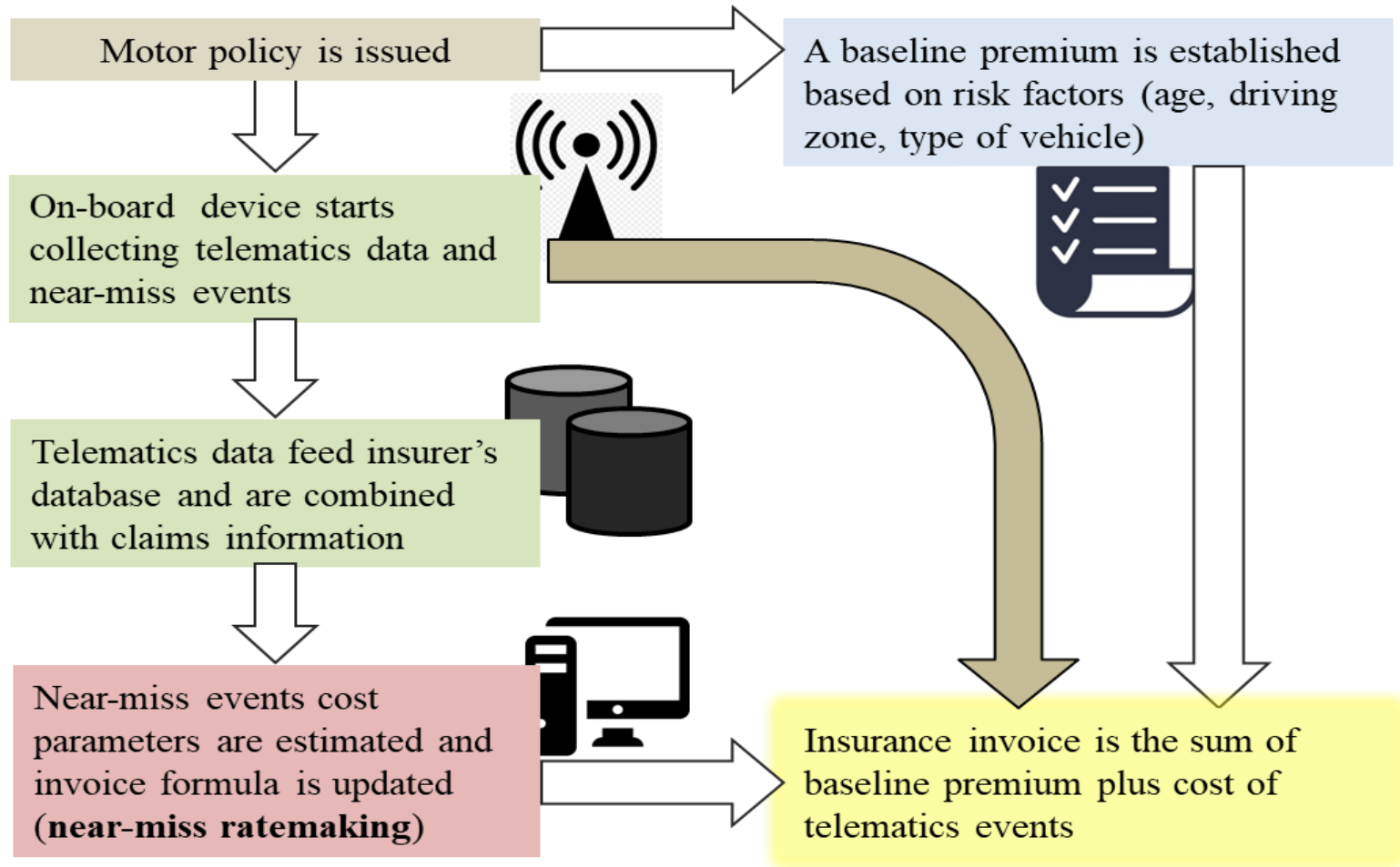
Problem: *(at fault near-misses?)*

- (Very recent...) **Events of excess speed**

The driver exceeds by more than 10% the legal speed limit during one trip.

# Near-miss telematics

motor insurance pricing



Guillen et al. (2021)

# Notation and Poisson model specification

- $Y_i$  number of claims at fault policy  $i$ ,  
 $i = 1, \dots, n$
- $T_i$  risk exposure, offset for policy  $i$
- $x_i, E_i$  vectors of ratemaking factors  
(traditional, telematics,  $E_i$  stands for “events”)

$$\begin{aligned} E(Y_i | x_i, E_i) &= T_i \exp(x_i' \beta + E_i' \alpha) \\ &= T_i \exp(x_i' \beta) \exp(E_i' \alpha) \end{aligned}$$

# Transforming the log-link into an additive structure

- We do not consider distance driven ( $T_{it} = 1$ ). Coefficients,  $\alpha$ , are percent increases due to each near-miss event of the base rate  $P_{i \text{ base}}$ . An approximate linear rate that would penalize each additional near-miss is:

$$\bar{C} \exp(x'_i \beta) \exp(E'_{it} \alpha) = P_{i \text{ base}} \exp(E'_{it} \alpha) \cong P_{i \text{ base}} (1 + E'_{it} \alpha) \leq P_{i \text{ base}} + E'_{it} \alpha_{max},$$

$$\text{where } \alpha_{max} = \max_{1 \leq i \leq n} (\alpha P_{i \text{ base}}).$$

- Note that  $\alpha_{max}$  depends on the maximum value of  $P_{i \text{ base}}$ . In practice, in order to determine  $\alpha_{max}$  a reasonable threshold for  $P_{i \text{ base}}$  could be, for example, three times the average of  $P_{i \text{ base}}$ .

# Aproximate additive structure & linearising exposure to risk

- The following approximation for the weekly premium that would penalize each additional near-miss ( $E_{it}$ ) and each additional unit of distance ( $T_{it} > 0$ ) is:

$$\bar{C} T_{it} \exp(x_i' \beta) \exp(E_{it}' \alpha) = P_{i \text{ base}} T_{it} \exp(E_{it}' \alpha) \cong P_{i \text{ base}} (1 + E_{it}' \alpha + \ln(T_{it})) \leq$$

$$P_{i \text{ base}} + E_{it}' \alpha p_{max} + p_{max} \ln(T_{it}),$$

$$\text{where } p_{max} = \max_{1 \leq i \leq n} (P_{i \text{ base}}), \alpha_{max} = \alpha p_{max} \cdot$$

# Data

- Anonymous telematics information from 641 drivers
- Collected between 30th week of 2016 until the 30th week of 2019 & past claims records
- Southern Europe
- **Weekly data** 7,570 vehicle-week observations in the telematics data set

TABLE 2 Descriptive statistics in the telematics and claims data sets

Variable	Mean	SD	Minimum	Maximum
EBrak1	1.8629	6.3567	0	93
EBrak2	0.6764	2.7104	0	33
EBrak3	0.1703	1.1368	0	29
EBrak	2.7095	8.6934	0	119
EAclr1	1.3931	7.1128	0	202
EAclr2	0.1180	0.7488	0	20
EAclr3	0.1655	1.1916	0	30
EAclr	1.6766	7.9614	0	219
EPhone	16.0008	77.7907	0	4150
DistThous	0.1523	0.1865	0.0010	2.7230
EngineCapacity	1.8383	0.7328	0.4250	6.2550
NumT	0.0764	0.4352	0	6
ExpoT	287.3532	53.5283	52.2857	545.8571

# Near-miss telematics

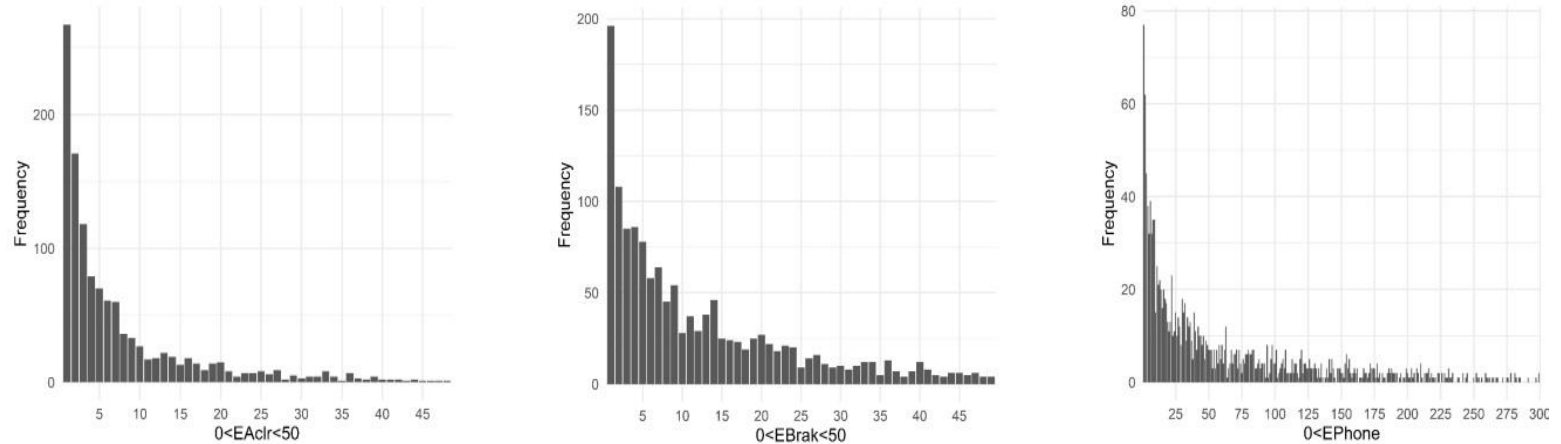


Figure 1, 2 and 3 show the histogram of EBrak, EAclr and EPhone. Due to the large frequency of zeroes we decided to remove them from the graphs, therefore only positive observations are represented. The data present a long right tail, so we also decided to limit the representation up to a maximum value, specifically 50 for EAclr and EBrak, and 300 for EPhone. Note that EAclr has 83.66% of zeroes, and 0.62% are equal or greater than 50. EBrak has 80.91% of zeroes, and 0.82% are equal or greater than 50, and finally EPhone has 79.78% of zeroes and 0.65% are higher than 300.



# Claims frequency using near-miss events as covariates

**Table 3. Parameter estimates of the Poisson model of the weekly rate of at fault claims for the telematics and claims data set**

Parameter	Estimate	Standard Error	p-value
Intercept	-8.0637	0.0673	<.0001
EAclr1	-0.0825	0.0265	0.0019
EAclr2	0.3069	0.1277	0.0162
EAclr3	0.0095	0.0390	0.8072
EBrak1	0.0268	0.0086	0.0018
EBrak2	-0.4966	0.0770	<.0001
EBrak3	0.0984	0.0336	0.0034
EPhon	0.0004	0.0002	0.0776
EngineCapacity	0.3644	0.0287	<.0001

The AIC equals 7345.00 and the BIC equals 7407.39. The pseudo-R<sup>2</sup> equals 21.83%.

# Claims frequency using near-miss events as covariates

**Table 3. Parameter estimates of the Poisson model of the weekly rate of at fault claims for the telematics and claims data set**

Parameter	Estimate	Standard Error	p-value
Intercept	-8.0637	0.0673	<.0001
EAc1r1	-0.0825	0.0265	0.0019
EAc1r2	0.3069	0.1277	0.0162
EAc1r3	0.0095	0.0390	0.8072
EBrak1	0.0268	0.0086	0.0018
EBrak2	-0.4966	0.0770	<.0001
EBrak3	0.0984	0.0336	0.0034
EPhon	0.0004	0.0002	0.0776
EngineCapacity	0.3644	0.0287	<.0001

The AIC equals 7345.00 and the BIC equals 7407.39. The pseudo-R<sup>2</sup> equals 21.83%.

# Claims frequency using near-miss events as covariates

**Table 3. Parameter estimates of the Poisson model of the weekly rate of at fault claims for the telematics and claims data set**

Parameter	Estimate	Standard Error	p-value
Intercept	-8.0637	0.0673	<.0001
EAc1r1	-0.0825	0.0265	0.0019
EAc1r2	0.3069	0.1277	0.0162
EAc1r3	0.0095	0.0390	0.8072
EBrak1	0.0268	0.0086	0.0018
EBrak2	-0.4966	0.0770	<.0001
EBrak3	0.0984	0.0336	0.0034
EPhon	0.0004	0.0002	0.0776
EngineCapacity	0.3644	0.0287	<.0001

The AIC equals 7345.00 and the BIC equals 7407.39. The pseudo-R<sup>2</sup> equals 21.83%.

# Near-miss telematics ratemaking

- Basic rate plus additional cost of near misses.

**Table 1. Weekly breakdown of a total bill per week. Pure premium in motor insurance as a function of near-miss events for a driver of a car with engine capacity 1,769 cc. Basic weekly rate 1.95 Eur.**

Week	Distance driven (km)	Number of near-miss brakes (a)	Number of near-miss accelerations (b)	Minutes of smart phone use (c)	Cost of near-misses (Eur) (d)	Bill per week (Eur) (e)
1	30	0	0	0	0,00	1,95
2	73	0	0	2	0,37	2,32
3	104	2	2	2	6,59	8,54
4	260	6	2	1	9,40	11,35
5	705	19	4	21	27,51	29,46

Total bill for five weeks: 53.61 Eur.

(e)=1.95+(d)

(d)=0.75(a)+2.36(b)+0.18(c)

# Near-miss telematics ratemaking

- Basic rate with a reward for safe driving and additional charge for near misses.

**Table 2. Weekly bill of pure premium in motor insurance as a function of near-miss events for a driver of a car with engine capacity 1,769 cc). Basic weekly rate (6.66 Eur) minus discounts for safe driving, or plus penalizations for near misses.**

Week	Distance driven (km)	Number of near-miss brakes (a)	Number of near-miss acceleration (b)	Minutes of smart phone use (c)	Cost of near-misses (Eur) (d)	Total weekly bill (Eur) (e)
1	30	0	0	0	-5.65	1.01
2	73	0	0	2	-5.29	1.37
3	104	2	2	2	0.93	7.59
4	260	6	2	1	9.00	15.66
5	705	19	4	21	54.94	61.60

Total bill for five weeks: 87.23 Eur

(e)=6.66+(d)

(d)=if ((a)>2, 1.5(a), -0.75(1-(a)),)+if ((b)>2, 4.71(b), -2.36(2-(b)))+if ((c)>2, 0.36(c), -0.18(1-(c)))

# Case study I: Take aways

- Driver **pays per risky-events/ gets a discount for absence** of risky-events.
- We are **unable to say** from our empirical analysis whether drivers **adopting telematics schemes will in general change their behavior in the long term** as a consequence of the impact on the price of their usage-based insurance ratemaking.
- Near-miss ratemaking is **easily introduced**. After some weeks, an insurer can start pricing and re-adjust the formula to improve predictive performance and fairness.

→ First step towards **Pay-How-You-Drive (PHYD)** on a **Pay per trip** schemes! And ....**Pay-Where-You-Drive (PWYD)**.

# Journal of Risk and Insurance (2021)

Received: 5 September 2020 | Revised: 23 March 2021 | Accepted: 27 March 2021

DOI: 10.1111/jori.12340

ORIGINAL ARTICLE

Journal of Risk and Insurance

## Near-miss telematics in motor insurance

Montserrat Guillen<sup>1</sup>  | Jens Perch Nielsen<sup>2</sup>  |  
Ana M. Pérez-Marín<sup>1</sup> 

<sup>1</sup>Department of Econometrics,  
Riskcenter-IREA, Universitat de  
Barcelona, Barcelona, Spain

<sup>2</sup>Cass Business School, University of  
London, London, United Kingdom

### Correspondence

Montserrat Guillen, Department of  
Econometrics, Riskcenter-IREA,  
Universitat de Barcelona, Av. Diagonal,  
690, 08034 Barcelona, Spain.  
Email: mguillen@ub.edu

### Funding information

Ministerio de Ciencia e Innovación,  
Grant/Award Number: PID2019-  
105986GB-C21; Institució Catalana de  
Recerca i Estudis Avançats,  
Grant/Award Number: ICREA  
Acadèmia; Fundació BBVA,  
Grant/Award Number: Research in  
Big Data

### Abstract

We present a method to integrate telematics data in a pay-how-you-drive insurance pricing scheme that penalizes some near-miss events. We illustrate our method with a sample of drivers for whom information on near-miss events and claims frequency records are available. We discuss the implications for motor insurance ratemaking. Our pricing principle is to combine a baseline insurance premium with added extra charges for near-miss events indicating risky driving (or discounts) that can be updated on a weekly basis. This procedure provides an incentive for safe driving. In our real-case study illustration, hard-braking and acceleration events as well as smartphone use while driving increase the cost of insurance.

### KEYWORDS

claims frequency, dynamic ratemaking, Poisson model, pricing

# CASE STUDY II

- Pricing with near-misses
- Contextual data





Photo: Zachary DeBottis

# Driving context and hazardous patterns

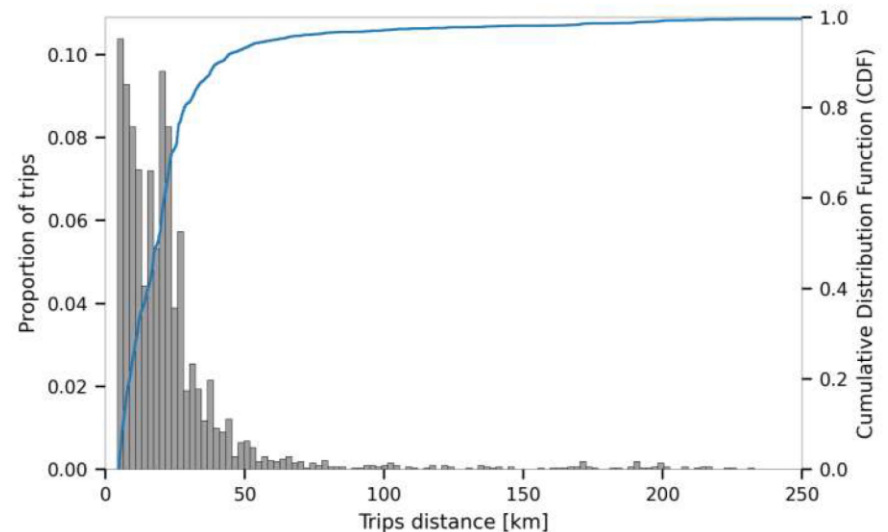
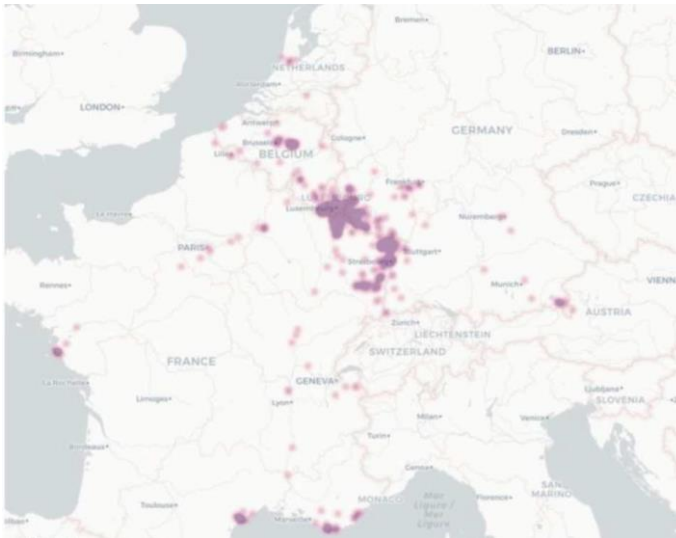
- i) Does the **driving context** carry predictive power for **measuring driving risk**?
- ii) What are the **most relevant contextual features** to evaluate risk exposure?
- iii) How **contextual factors** influence the **occurrence of risky-events (near-misses)** and hazardous patterns?

# The influence of the driving context

- Most commonly studied contextual factors categories:
  - Road environment
  - Road infrastructure and topology
  - Traffic conditions
  - Road signs
  - Weather and lighting conditions
- There is not specific research on how driving under several combinations of these factors influences exposure to dangerous events

# Data

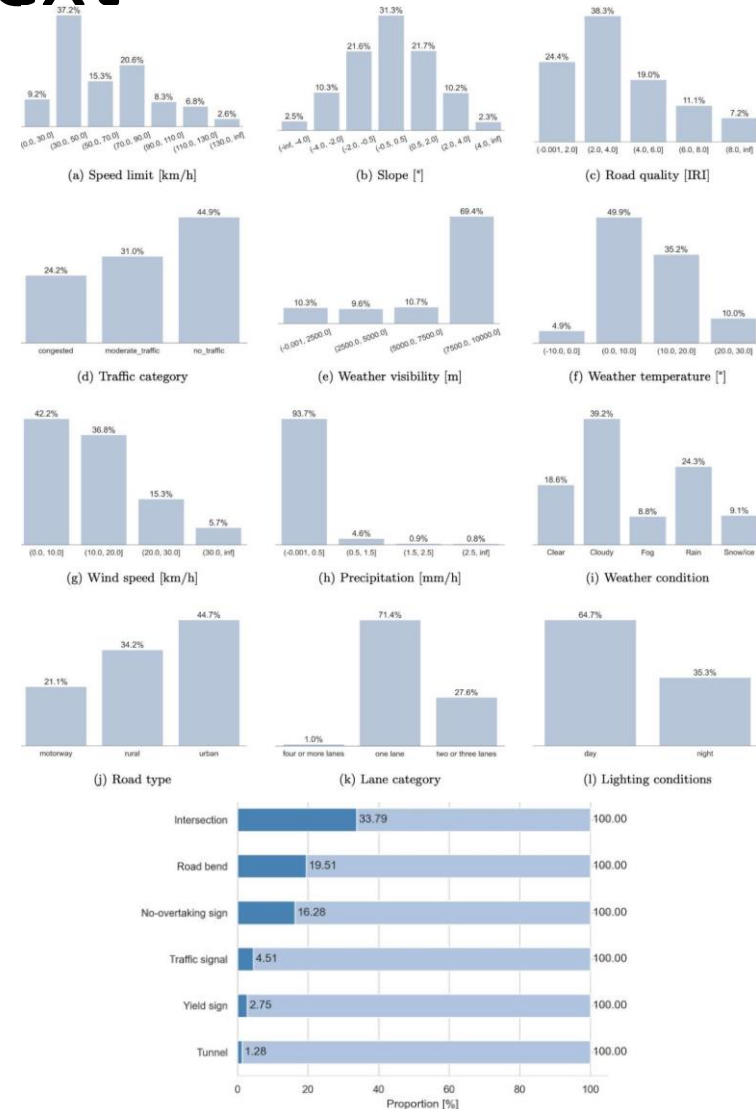
- Anonymous telematics information from 32 drivers
- Collected between July, 2021 and February, 2022
- Belgium, France, Germany, Luxembourg, Netherlands



3,220 trips with at least 5km  $\rightarrow$  77,859 km (24,9km/trip)

# Contextual attributes of the driving context

Contextual Category	Attribute
Environment	Speed limit [km/h]
	Road type
Road infrastructure and topology	Intersection (distance to intersection less than 30 m)
	Road quality (IRI)
	Lane category
	Road bend
	Road slope [°]
	Tunnel
Traffic conditions	Traffic category
Traffic signs	Presence of no-overtaking sign
	Presence of traffic lights
	Presence of yield sign
Weather and lighting information	Lighting conditions
	Precipitation [mm/h]
	Temperature [°C]
	Visibility [m]
	Weather conditions
	Wind speed [km/h]



# Target risky events

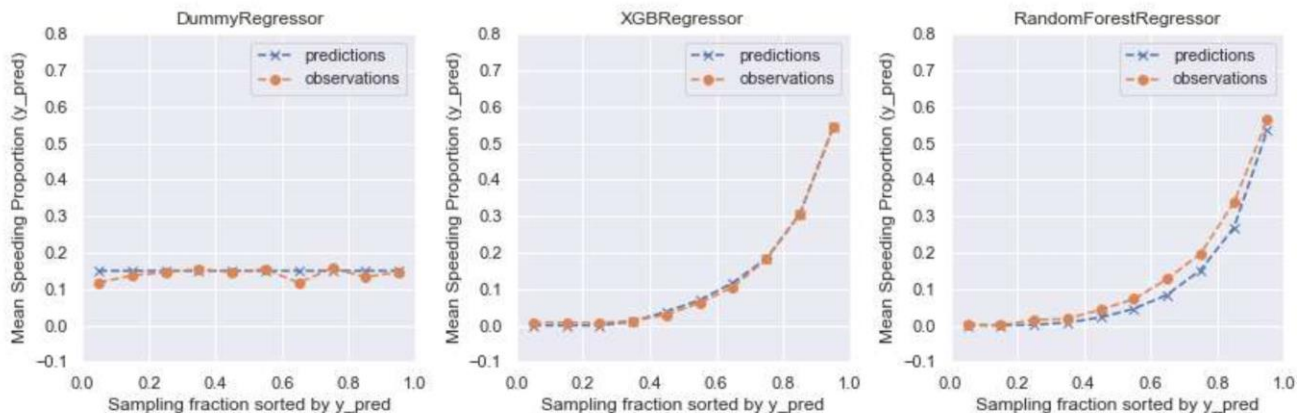
- **Cornering**: accelerations exceeding  $7.5 \text{ m/s}^2$  in curves
- **Harsh Acceleration**: accelerations exceeding  $6 \text{ m/s}^2$
- **Harsh Braking**: brakings with magnitudes less than  $-6 \text{ m/s}^2$
- **Phone Call**: % kilometres driven while making phone calls
- **Phone Unlocking**
- **Speeding**: % kilometres driven exceeding the legal speed limit

# Modelling techniques in driving risk assessment

- What models can be used?
  - Generalized Linear Models (GLM)
  - Machine Learning Techniques (ML):
    - Logistic Regression
    - Neural Networks
    - Decision Trees
- Trade-off → Predictive gains vs. interpretability
- Explainable Artificial Intelligence
  - Shapley Additive Explanations (SHAP)

# Modelling techniques employed

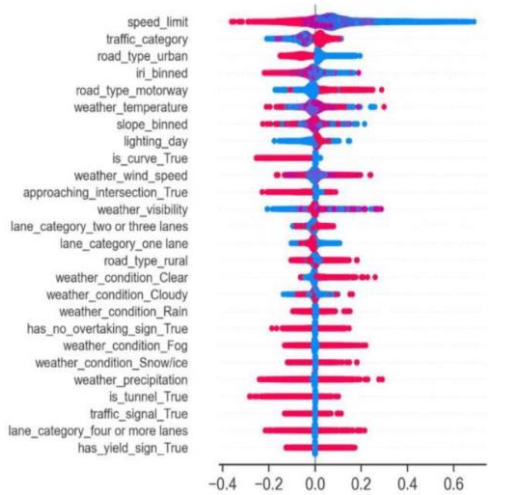
- Random Forest
- XGBoost
- Neural Networks
- GLM
- Model results:
  - XGBoost and Random Forest significantly outperform neural networks and GLM
  - **XGBoost is generally better than random forest**



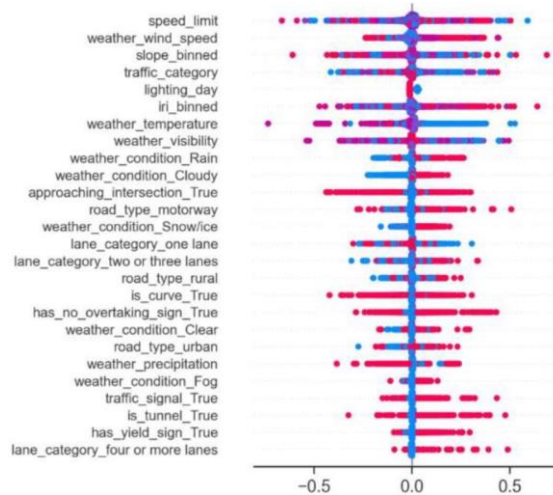


# Contextual feature importance for our six risky events: SHAP values

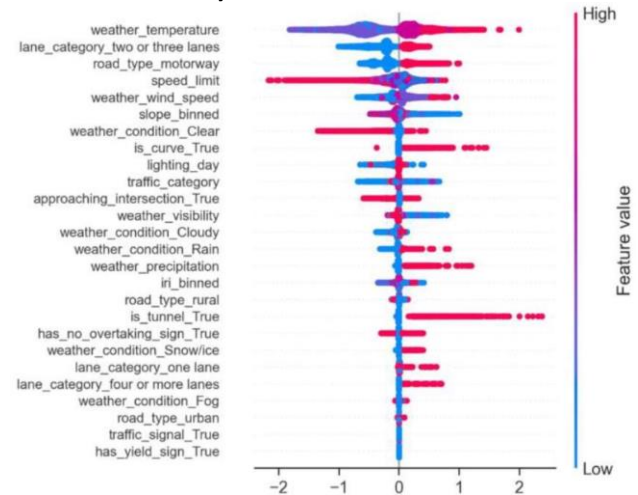
a) Speeding %



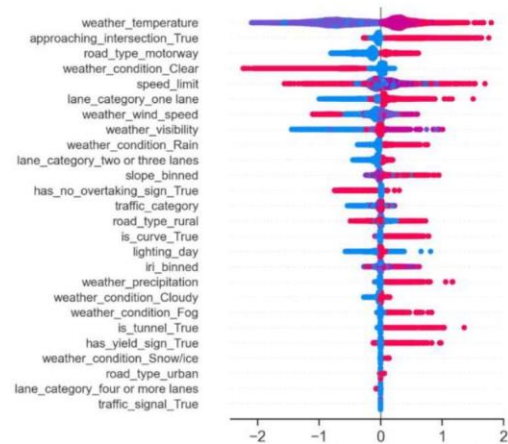
b) Phone Call %



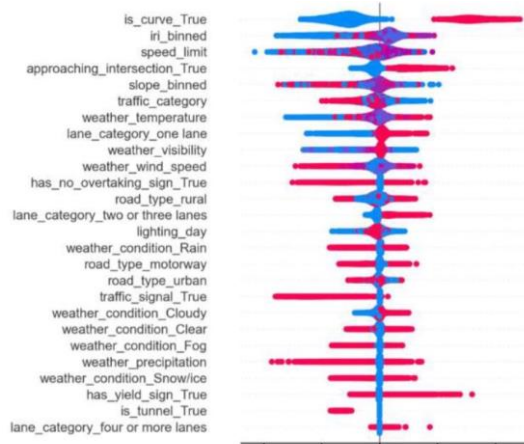
c) Harsh acceleration



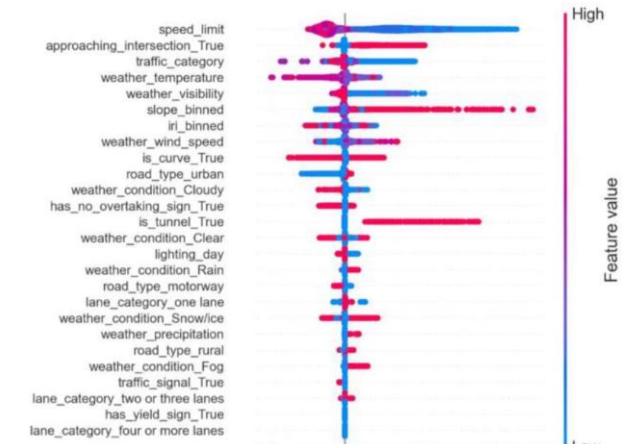
d) Harsh braking



e) Cornering

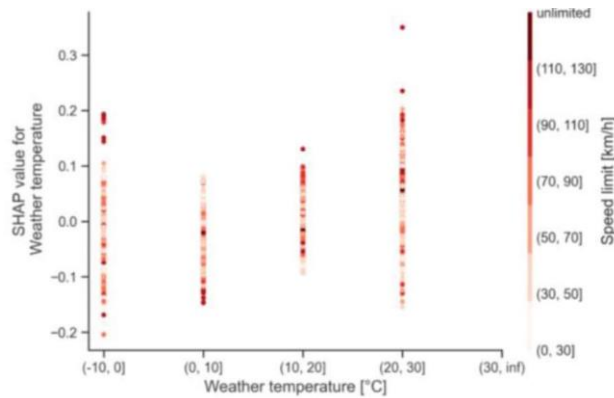


f) Phone Unlocking

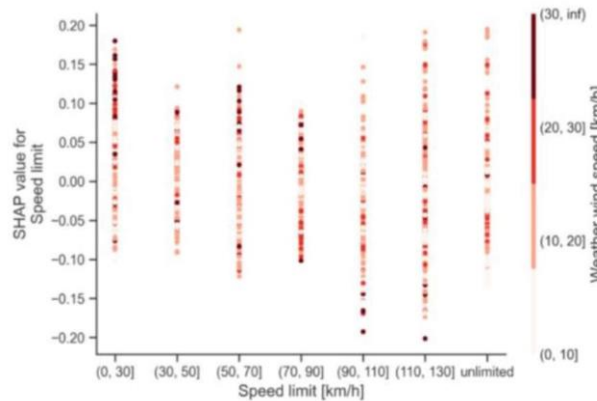


# SHAP dependence for top ranked contextual features

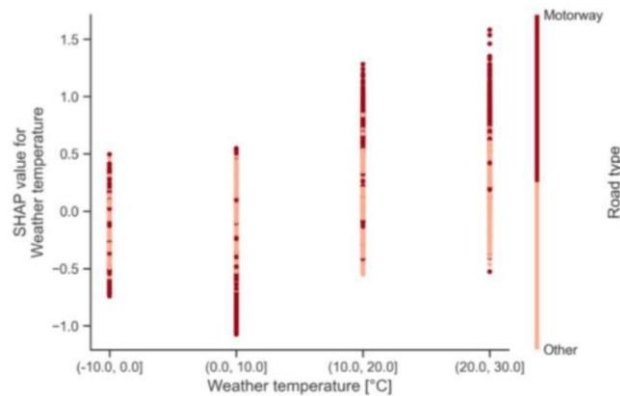
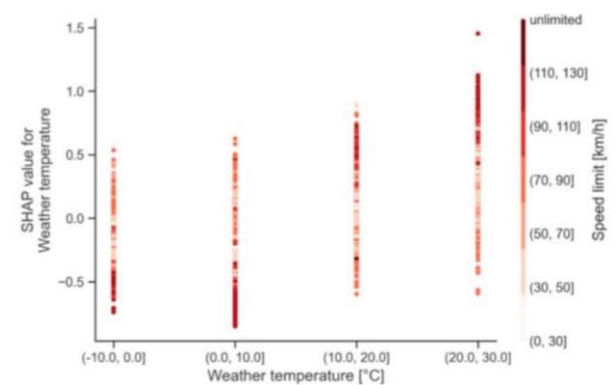
a) Speeding %



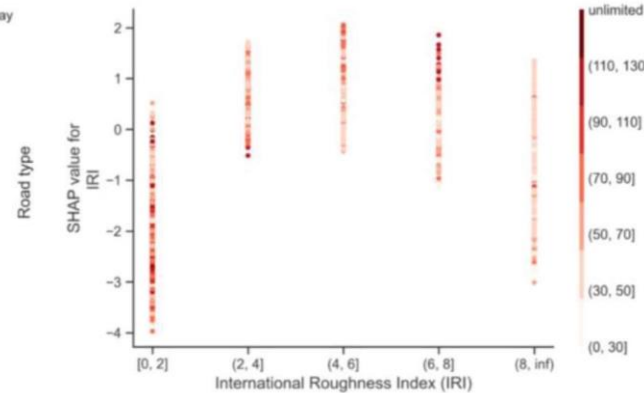
b) Phone Call %



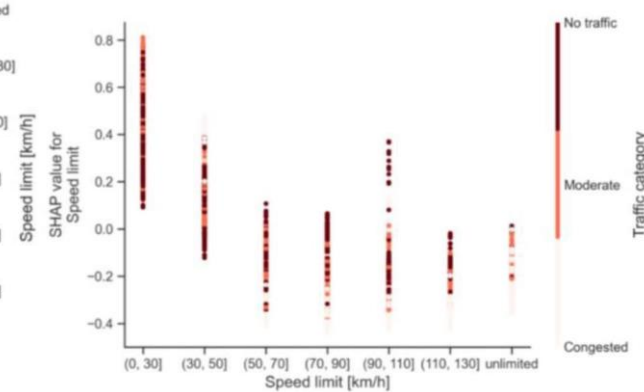
c) Harsh acceleration



d) Harsh braking

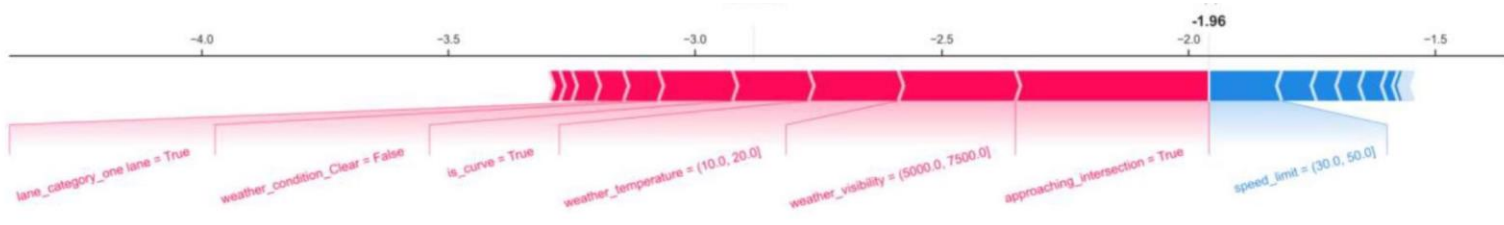


e) Cornering

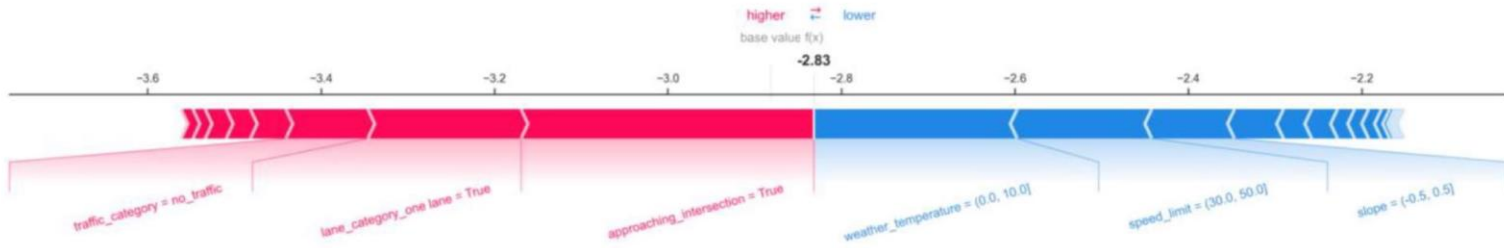


f) Phone Unlocking

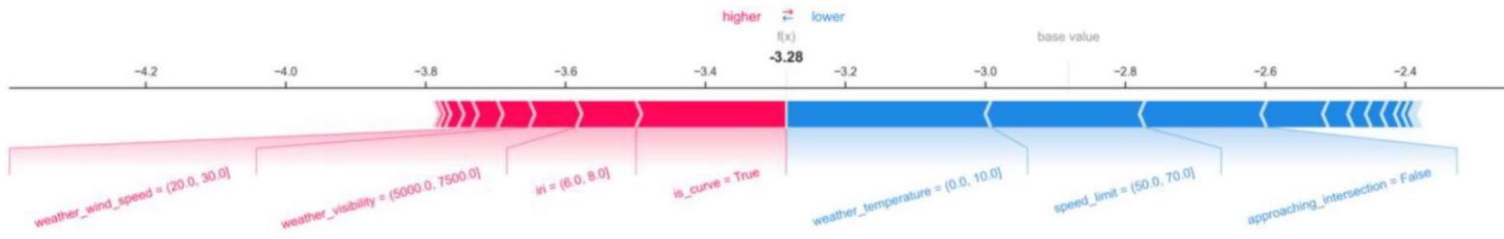
# How every contextual feature contributes to the model output?



(a) High-risk prediction



(b) Average-risk prediction



(c) Low-risk prediction

# Case study II: Take aways

- The **most powerful predicting features** are :
  - Speed limits
  - Weather temperatures
  - Wind speed
  - Traffic conditions
  - Road slope
- Moreover, **many combinations of contextual features are strongly associated with risky events**

→ First step towards **Pay-Where-You-Drive (PWYD)** schemes.

# Accident Analysis and Prevention (submitted)

Joint with **Leandro Masello**, German Castignani, Barry Sheehan and Finbarr Murphy

Using contextual data to predict risky driving events: A novel methodology from Explainable Artificial Intelligence

## Abstract

Usage-based insurance has allowed insurers to dynamically tailor insurance premiums by understanding when and how safe policyholders drive. However, telematics information can also be used to understand the driving contexts experienced by the driver within each trip (e.g., road types, weather, traffic). Since different combinations of these conditions affect the exposure to accidents, this understanding introduces predictive opportunities in driving risk assessment. This paper investigates the relationships between driving context combinations and risk using a naturalistic driving dataset of 77,859 km. In particular, XGBoost and Random Forests are used to determine the predictive significance of driving contexts for near-misses, speeding and distraction events. Moreover, the most important contextual factors in predicting these risky events are identified and ranked through Shapley Additive Explanations. The results show that the driving context has significant power in predicting driving risk. Speed limit, weather temperature, wind speed, traffic conditions and road slope appear in the top ten most relevant features for most risky events. Low-speed limits increase the predicted frequency of speeding and phone unlocking events, whereas high-speed limits decrease harsh accelerations. Low temperatures decrease the expected frequency of harsh manoeuvres, and precipitations increase harsh acceleration, harsh braking, and distraction events. Furthermore, road slope, intersections and pavement quality are the most critical factors among road layout attributes. The methodology presented in this study aims to support road safety stakeholders and insurers by providing insights to study the contextual risk factors that influence road accident frequency and driving risk.

**Keywords:** driving context; explainable AI; machine learning; risk assessment; usage-based insurance.

# Contents

1. Introduction
2. Methods
3. Case Studies
- 4. Conclusions & take-home**

# How will **motor insurance** **ratemaking** change?

- Consumers
  - Personalization
  - More interaction with insurers
- Manufacturers
  - Vehicles will be equipped with telematics and possibly vehicles provide a service (insurance included)
- Insurers
  - Products are more demanding 24/7
  - Data analysts are needed. **Preprocessing is crucial**
  - Communication to mass consumers of complex pricing
  - Prevention and service provision

# References

- Guillen, M., Pérez-Marin, A.M. and Nielsen, J.P. (2021) “Near-miss telematics in motor insurance” **Journal of Risk and Insurance**, <https://onlinelibrary.wiley.com/doi/epdf/10.1111/jori.12340>
- Guillen, M., Bermúdez, L.I. and Pitarque, A. (2021) “Joint Generalized Quantile and Conditional Tail Expectation Regression for Insurance Risk Analysis” **Insurance, Mathematics and Economics**, 99, 1-8.
- Guillen, M., Pérez-Marín, A.M. And Alcañiz (2021) “Percentile reference charts for speeding based on telematics information” **Accident Analysis and Prevention**, 2021, vol. 150, num. 105865
- Guillen, M., Nielsen, J.P., Pérez-Marín, A.M. and Elpidorou, V. (2020) “Can automobile insurance telematics predict the risk of near-miss events?” **North American Actuarial Journal**, 24, 1, 141-152.
- Pesantez-Narvaez, J. and Guillen M. (2020) “Weighted Logistic Regression to Improve Predictive Performance in Insurance” **Advances in Intelligent Systems and Computing**, 894, 22-34.
- Sun, S., Bi, J., Guillen, M. and Pérez-Marín, A. M. (2020) “Assessing driving risk using internet of vehicles data: an analysis based on generalized linear models” **Sensors**, 20(9), 2712
- Denuit, M., Guillen, M. and Trufin, J. (2019) “Multivariate credibility modeling for usage-based motor insurance pricing with behavioural data” **Annals of Actuarial Science**, 13, 2, 378-399.
- Pérez-Marín, A.M. and Guillen, M. (2019) “The transition towards semi-autonomous vehicle insurance: the contribution of usage-based data”, **Accident Analysis and Prevention**, 123, 99-106.
- Pérez-Marín, A.M., Ayuso M.M. and Guillen, M. (2019) “Do young insured drivers slow down after suffering an accident?”, **Transportation Research Part F: Psychology and Behaviour**, 62, 690-699.
- Guillen, M., Nielsen, J.P., Ayuso, M. and Pérez-Marín, A.M. (2019) “The use of telematics devices to improve automobile insurance rates”, **Risk Analysis**, 39, 3, 662-672.
- Ayuso, M., Guillen, M. and Nielsen, J.P. (2019) “Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data”, **Transportation**, 46(3), 735-752.
- Boucher, J.P., Coté, S. and Guillen, M. (2018) “Exposure as duration and distance in telematics motor insurance using generalized additive models”, **Risks**, 5(4), 54.





[mguillen@ub.edu](mailto:mguillen@ub.edu)

# Motor insurance with telematics driving data

Montserrat Guillen

Seminar organized by ETH Zurich, RiskLab

September 16, 2022