Pricing motor insurance with telematics data

Montserrat Guillen

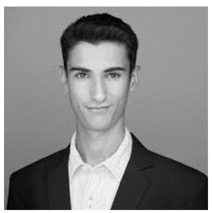
Chaire PARI, Novembre 15, 2023

Risk center RESEARCH

UNIVERSITAT DE BARCELONA

Ana M. Pérez-Marín

Jens P. Nielsen



Jan Reig

Lorena Rey

Anna Salazar

Iker Meneses

Giselle Aguer

Leandro Masello
& co-authors
University of Limerick
& Motion-S

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

ICREA

Contents

Contents

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

Samuele Errico Piccarini in Unsplash

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)

DATA IN MOTOR INSURANCE, EVERY 1’’

What is telematics data?

- Global Positioning Signal (GPS) –not always recorded-
- 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.

  PAY-**AS-**YOU-DRIVE-> PAY-**HOW**-YOU-DRIVE -> PAY-**WH**-YOU-DRIVE

- Telematics 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 beginings: PAYD, milleage 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-

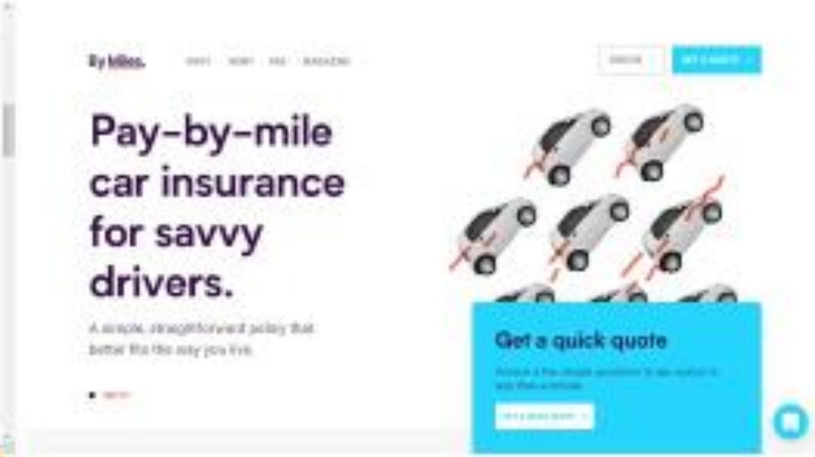
  Li et al. (2023) published ESWA –interpretable machine learning-

- 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) · **Risky events**
- 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,...)

Will telematics change ratemaking models in automobile insurance?

Companies selling motor insurance based on telematics around the world



PAY-AS-YOU-DRIVE PRICING = **BASE PREMIUM** + DISTANCE***COST per UNIT**

Telematics data: today in 2023

NASA in Unsplash

Contents

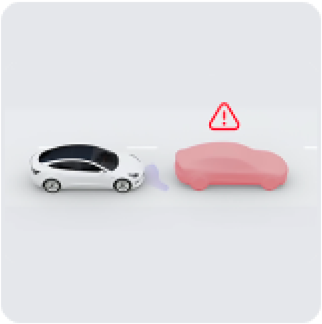
1. Introduction
2. Methods
3. Case Studies
4. Conclusions & take-home

Tesla’s Safety Score

Tesla's **Eight** Factors

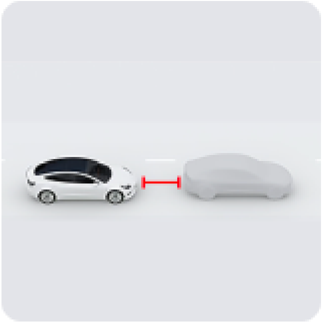
Forward Collision Warnings per 1,000 Miles*

*capped 117



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

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

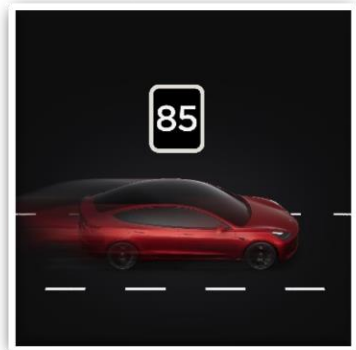


Unsafe Following [Prop. 1sec/3sec Speed >50mph (80Km/h)]

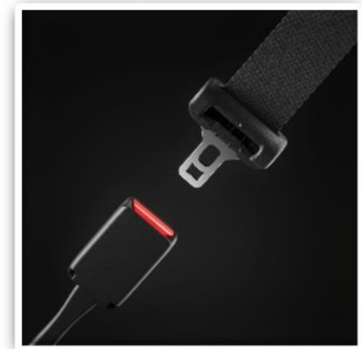
Forced Autopilot Disengagement [After 3 warnings of inatentive, no hands on the wheel

Tesla's **Eight** Factors

Late Night % Driving
10PM-4AM capped at approx. 30%



Excessive Speeding %
time spent driving in excess of 85 mph



Unbuckled Driving %
time pent driving above 10 mph without fastening the driver's seatbelt

https://www.tesla.com/support/safety-score

Tesla's Safety Score Version 1.0

Predicted Collision Frequency (PCF) =

$$0.68 \times 1.01^{\text{Forward Collision Warnings per 1,000 Miles}} \times 1.13^{\text{Hard Braking}} \times 1.02^{\text{Aggressive Turning}} \times 1.00^{\text{Unsafe Following Time}} \times 1.32^{\text{Autopilot Disengagement}}$$

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 x PCF

Tesla's Safety Score Version 1.2

Predicted Collision Frequency (PCF) =

0.42 x
$1.01^{\text{Forward Collision Warnings per 1,000 Miles}}$ x
$1.10^{\text{Hard Braking}}$ x
$1.00^{\text{Aggressive Turning}}$ x
$1.00^{\text{Unsafe Following Time}}$ x
$1.12^{\text{Autopilot Disengagement}}$ x
$\mathbf{1.03^{\text{Night driving}}}$

The current formula was derived based on statistical modeling using 8 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 = 112.31 - 29.33 x PCF

Tesla's Safety Score Version 2.0

Predicted Collision Frequency (PCF) =

$$0.83 \times$$
$$1.01^{\text{Forward Collision Warnings per 1,000 Miles}} \times$$
$$1.16^{\text{Hard Braking}} \times$$
$$1.01^{\text{Aggressive Turning}} \times$$
$$1.00^{\text{Unsafe Following Time}} \times$$
$$1.41^{\text{Autopilot Disengagement}} \times$$
$$\mathbf{1.05}^{\textbf{Night Driving}} \times$$
$$\mathbf{1.01}^{\textbf{Excessive Speeding}} \times$$
$$\mathbf{1.01}^{\textbf{Unbuckled Driving}}$$

The current formula was derived based on statistical modeling using 8 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 = 112.29 – 14.77 x PCF

Tesla's Safety Score 1.0 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}

PCF=1,13

Safety Score =

115.38 – 22.53 x PCF

Tesla's Safety Score 1.2 in log link

Predicted Collision Frequency (PCF) = exp{ -0.377 +

0.043 Forward Collision Warnings per 1,000 Miles +

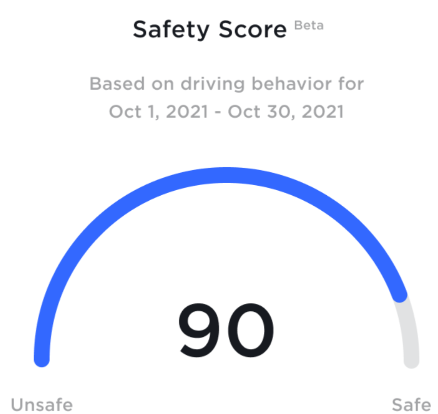
0.004 Hard Braking +

0.001 Aggressive Turning +

0.001 Unsafe Following Time +

0.047 Autopilot Disengagement +

0.014 Percent Night Driving}



Safety Score =

= 112.31 - 29.33 x PCF

PCF=0,76

Yearly Accident frequency to Safety Score 1.0 (aprox. Equivalence)

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
    - **Speed limits** / independent of posted limit
    - --- Performance relative to other drivers.

New 2023

- What type of collisions? unknown

An overview of methods when claims information is available

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, Shap values

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

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

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

Guillen et al. (2021) & Gao, Meng, Wüthrich (2022)

Poisson deviance loss

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

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

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

$\mathcal{T}$ 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 $\varphi()$ 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:

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

$$\mathrm{T_i} \exp(\mathrm{x_i'}\beta) \exp(\mathrm{z_i}'\alpha)$$

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

$$\mathrm{E}(\mathrm{Y_i}|\mathrm{x_i}, \mathrm{z_i}, \mathrm{T_i}) =$$

$$\mathrm{T_i} \exp(\mathrm{x_i'}\beta) \exp\left(\alpha_0 + \alpha_1 \bar{\psi}_i\right)$$

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

ML methods and interpretability

Response outcome is called Y_i, x_i, z_i vectors of ratemaking factors (traditional x_i , telematics z_i)

- Machine learning (GLM and **enhanced GLM**, lower level than CANN)
- Trees
- Random forests
- **Neural Networks** has more layers that allow it to learn more complex relationships between the inputs and outputs
- **Light Gradient Boost** combines simple models (weak learners) to increase prediction accuracy

Panel binary model specification (timeframe: weekly data)

- $\mathrm{Y}_{\mathrm{i}t}$ 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 , (?)
- $\mathrm{x}_{\mathrm{i}}, \mathrm{z}_{\mathrm{it}}$ vectors of ratemaking factors (traditional x_{i}, telematics/context/dynamic z_{it})
- We assume a panel structure where Y_{it} are independent across all policy holders. If there is independence over time:

$$\mathrm{E}(\mathrm{Y}_{\mathrm{i}t} | \mathrm{x}_{\mathrm{i}}, \mathrm{z}_{\mathrm{it}}, \mathrm{T}_{\mathrm{it}}) = \mu(\mathrm{x}_{\mathrm{i}}, \mathrm{z}_{\mathrm{it}}, \mathrm{T}_{\mathrm{it}})$$
$$= Prob(\mathrm{Y}_{\mathrm{i}t} = 1 | \mathrm{x}_{\mathrm{i}}, \mathrm{z}_{\mathrm{it}}, \mathrm{T}_{\mathrm{it}}) = \mathrm{p}_{\mathrm{it}}$$

Work in progress at RISKcenter UB

Panel binary model specification (timeframe: weekly data)

- Consider all information to (t-1), Ξ_{t-1}:

$$Prob(\mathrm{Y}_{\mathrm{i}t} = 1 | \mathrm{x_i}, \mathrm{z_{it}}, \mathrm{T_{it}}, \Xi_{t-1}) = \mathrm{p_{it}}$$

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

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

Work in progress at RISKcenter UB

Expression for weekly premium calculation

- Linear approximation:

$$\sum_{k=1}^{K} \alpha_k \, x_{ki} + \gamma lnD_{it} + (1-\gamma)\left[\sum_{j=1}^{J} \beta_{2j} z_{jit} + \sum_{l=1}^{L} \beta_{3j} z_{li(t-1)}\right]$$

Expression for weekly premium calculation

- Linear approximation:

$$\sum_{k=1}^{K} \alpha_k \, x_{ki} + \gamma ln D_{it} + (1-\gamma)\left[\sum_{j=1}^{J} \beta_{2j} z_{jit} + \sum_{l=1}^{L} \beta_{3j} z_{li(t-1)}\right]$$

Fixed over time: depends of classical covariates

Expression for weekly premium calculation

- Linear approximation:

$$\sum_{k=1}^{K} \alpha_k \, x_{ki} + \gamma lnD_{it} + (1-\gamma)\left[\sum_{j=1}^{J} \beta_{2j} z_{jit} + \sum_{l=1}^{L} \beta_{3j} z_{li(t-1)}\right]$$

Combination (γ) of log-distance driven, $(1-\gamma)$ other dynamic (current and previous)

Expression for weekly premium calculation

- Linear approximation:

$$\sum_{k=1}^{K} \alpha_k \, x_{ki} + \gamma lnD_{it} + (1-\gamma)\left[\sum_{j=1}^{J} \beta_{2j} z_{jit} + \sum_{l=1}^{L} \beta_{3j} z_{li(t-1)}\right]$$

Depends on J current week telematics and
K previous week telematics

Expression for weekly premium calculation

- Linear approximation:

$$\sum_{k=1}^{K} \alpha_k \, x_{ki} + \gamma ln D_{it} + (1-\gamma)\left[\sum_{j=1}^{J} \beta_{2j} z_{jit} + \sum_{l=1}^{L} \beta_{3j} z_{li(t-1)}\right]$$

May include context data (weather, road condition, traffic congestions)

Contents

CASE STUDY I

- Pricing with near-misses
- Weekly data
- Interpretable ML
- Scoring risky drivers

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

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

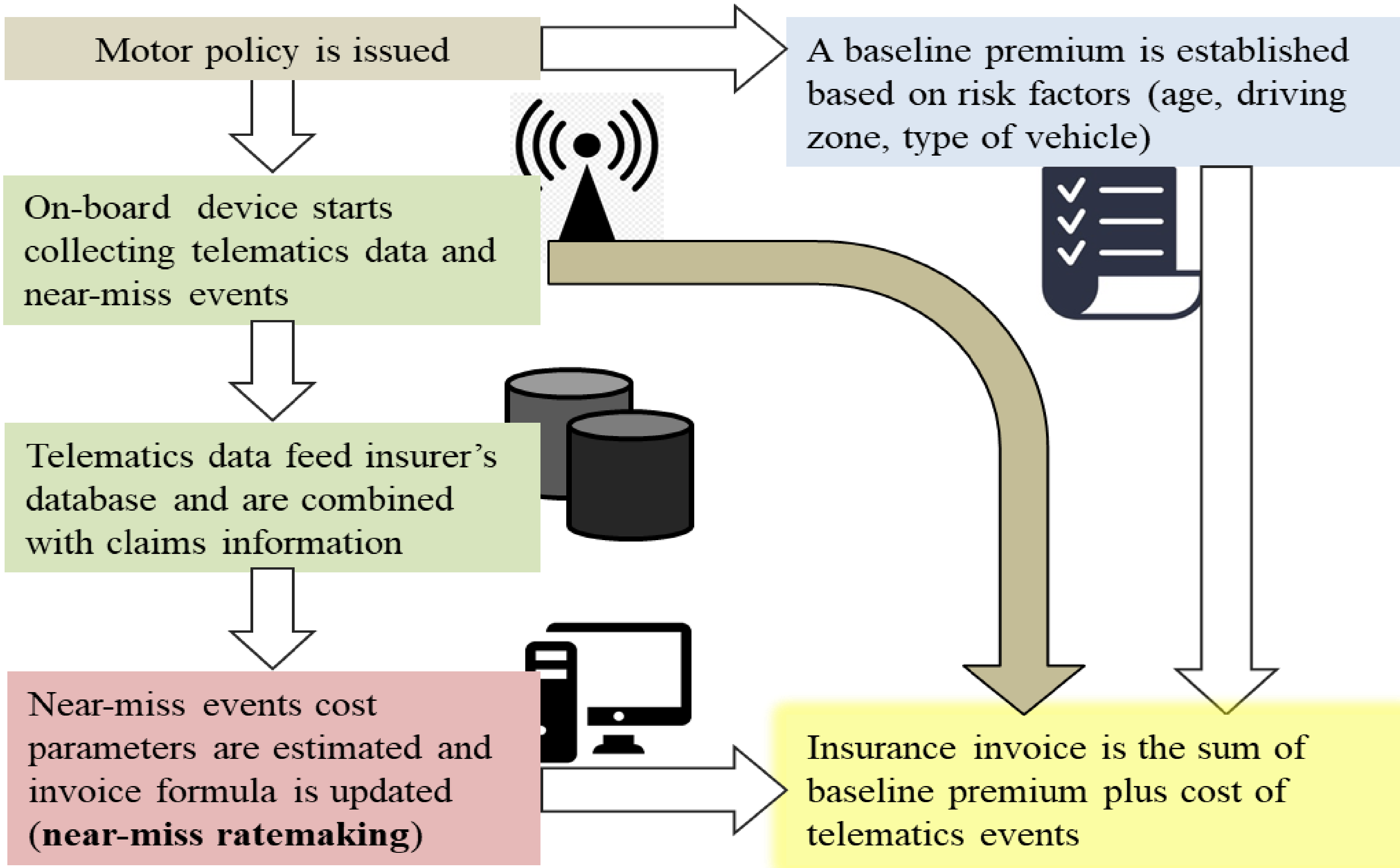
Problem: *(at fault near-misses?)*

- **(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)

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:

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

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

Guillen et al. (2022) in progress

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).

CASE STUDY II

- Pricing with near-misses
- Weekly data
- Interpretable ML
- Scoring risky drivers

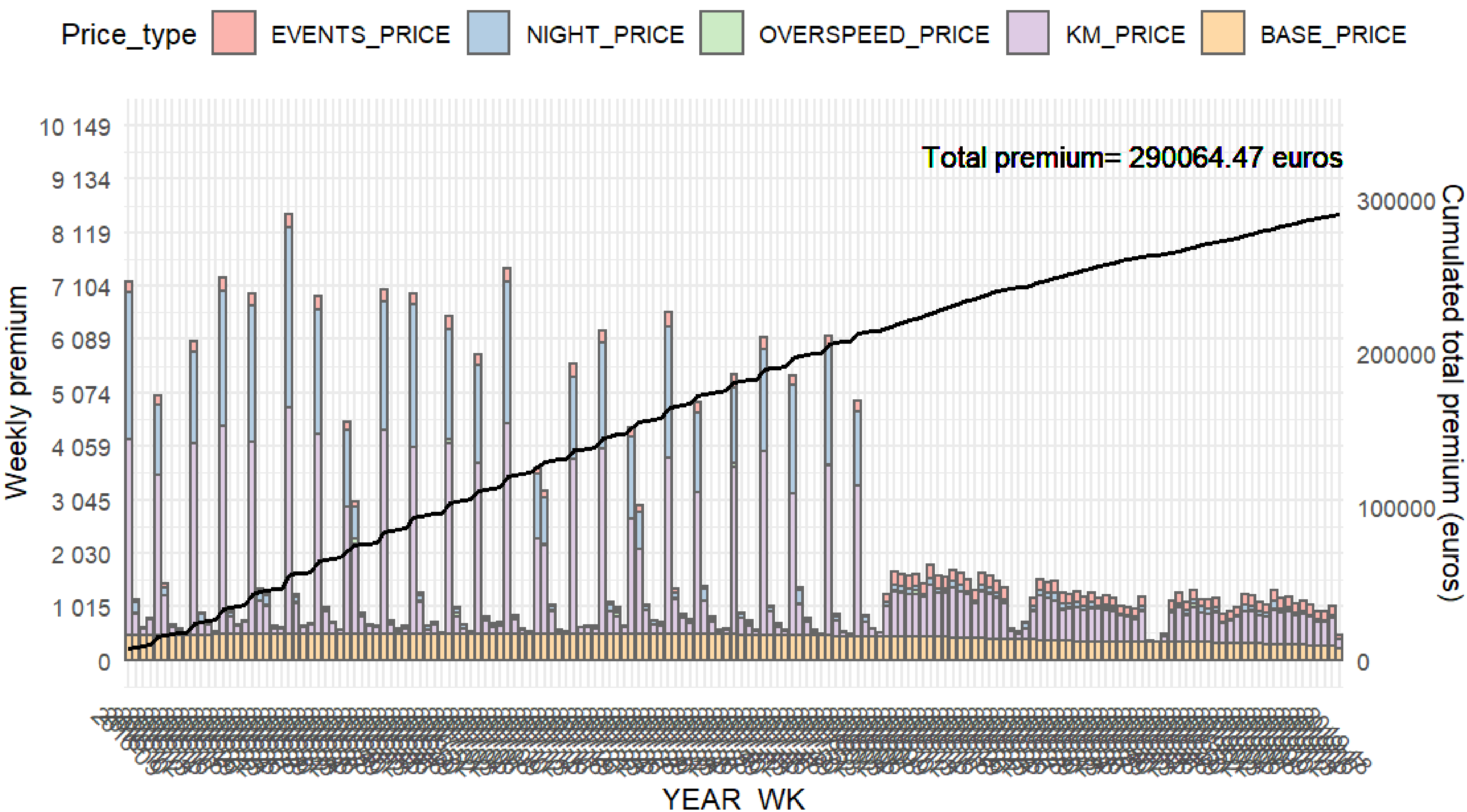
Cris Liverani in Unsplash

Data

- Anonymous data were provided by a Spanish insurer that commercializes pay-as-you drive-insurance since 2009.
- Specifically, our data contain **19,214** drivers observed in Spain from the 9th week of 2016 to the 18th week of 2019.
- We only have reliable weekly information from the 9th week of 2018 onwards, and 8 additional weeks were finally not considered valid in the analysis as there were very few observations, so they were eliminated during the data pre-processing.
- **930** claims at fault in the period of analysis.

Data per week | Data per policy | Plot per week | Plot Premiums/Claims

Total premiums per week



Driving claims weekly

i) Does the weekly information carry predictive power for measuring driving risk?
ii) What are the most relevant challenges to deal with dynamics? Is autoregressive a good idea?
iii) How do weekly premiums work?

Target risky Events & lagged telematics information

- Current week log-distance in 1,000 km
- Current week log-night distance in 1,000 km
- Percent urban driving distance with respect to total

- Previous week log-distance in 1,000 km
- Speeding: number of trips exceeding the legal speed limit in urban area the previous week

Autoregressive structure results

Table 6. Poisson Regression Models: Model 5a (all non-telematics, distance, urban speed events), Model 5b (all non-telematics, distance, percentage of urban driving), Model 6 (all non-telematics and lagged telematics variables), Model 7 (all non-telematics, lagged total distance, lagged percentage of urban driving and current total distance travelled at night) and Model 8 all non-telematics, current and lagged total distance, lagged percentage of urban driving and current total distance travelled at night) in telematics weekly data set, Spain 2019.

Variable	MODEL 5a		MODEL 5b		MODEL 6		MODEL 7		MODEL 8	
Intercept	-6.4387	<.0001	-6.3475	<.0001	-6.3493	<.0001	-6.2995	<.0001	-6.2474	<.0001
Vehicle_power	0.0016	0.1573	0.0020	0.0719	0.0020	0.0718	0.0020	0.0763	0.0019	0.0889
Gender	0.1538	0.0251	0.1502	0.0286	0.1508	0.0284	0.1379	0.0451	0.1374	0.0459
Age	-0.0161	0.0286	-0.0166	0.0236	-0.0167	0.0240	-0.0144	0.0505	-0.0146	0.0482
Ln(Total distance drivenKM)									0.0721	0.0914
Ln(Total distance drivenKM)_lag	0.1053	0.0062	0.3751	<.0001	0.3763	<.0001	0.3602	<.0001	0.3234	<.0001
Speed_event_urban_lag	0.0469	0.0027								
Perc_urban_lag			0.0134	<.0001	0.0134	<.0001	0.0132	<.0001	0.0134	<.0001
ln_km_nightMK							0.0103	0.0267	0.0086	0.0673
ln_km_nightMK_lag					-0.0005	0.9147				
AIC	14394.4084		14349.6923		14351.6808		14346.7992		14345.8938	

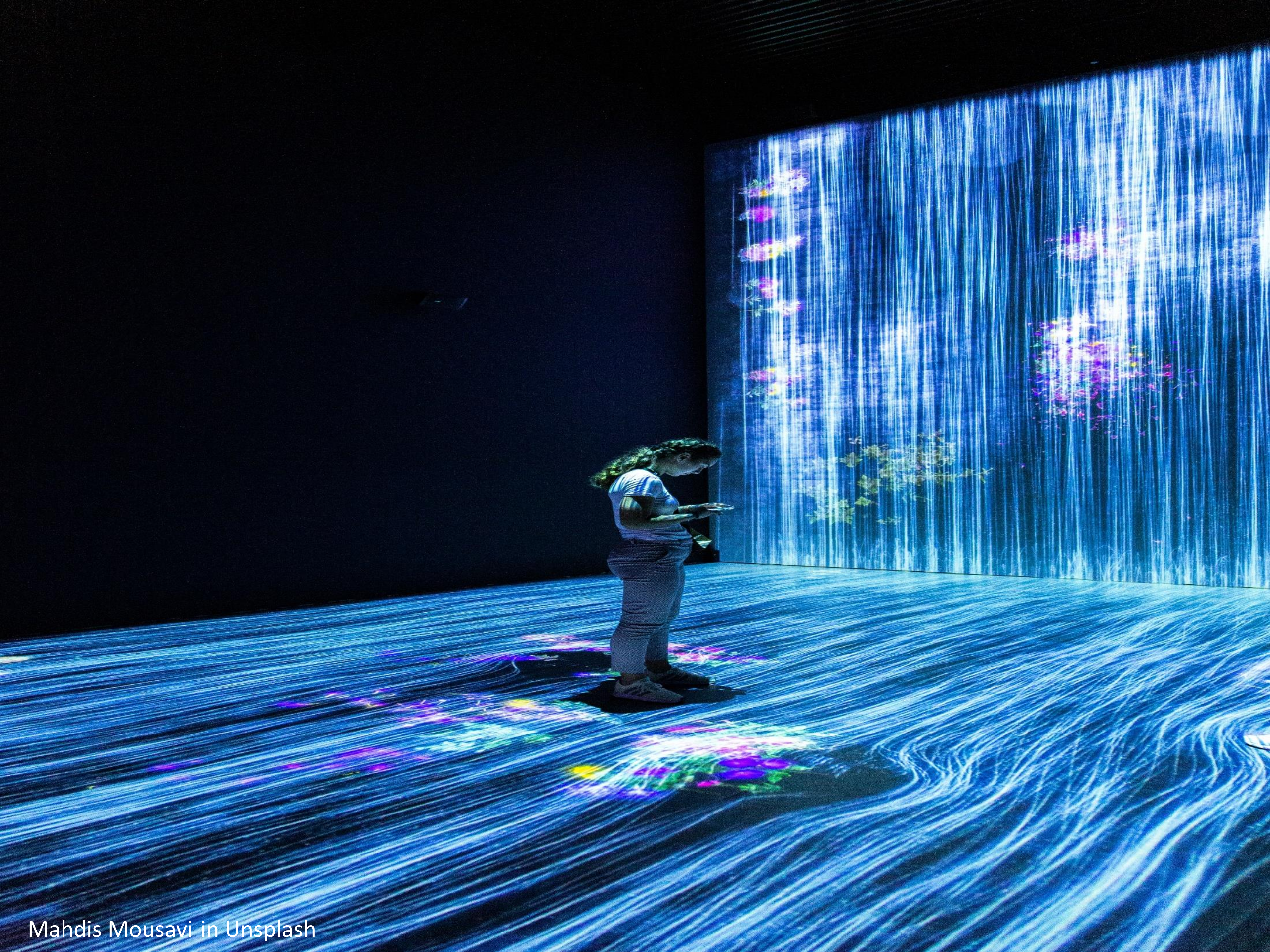
Case study II: Take aways

- The most powerful predicting features are :
  - Speed limits
  - Wind speed (not seen here)
  - Night time driving
  - Total kilometres driven
- Moreover, many combinations of contextual features are strongly associated with risky events

→ First step towards Pay-How-Where-You-Drive (PHWYD) schemes.

CASE STUDY III

- Pricing with near-misses
- Weekly data
- Interpretable ML
- Scoring risky drivers



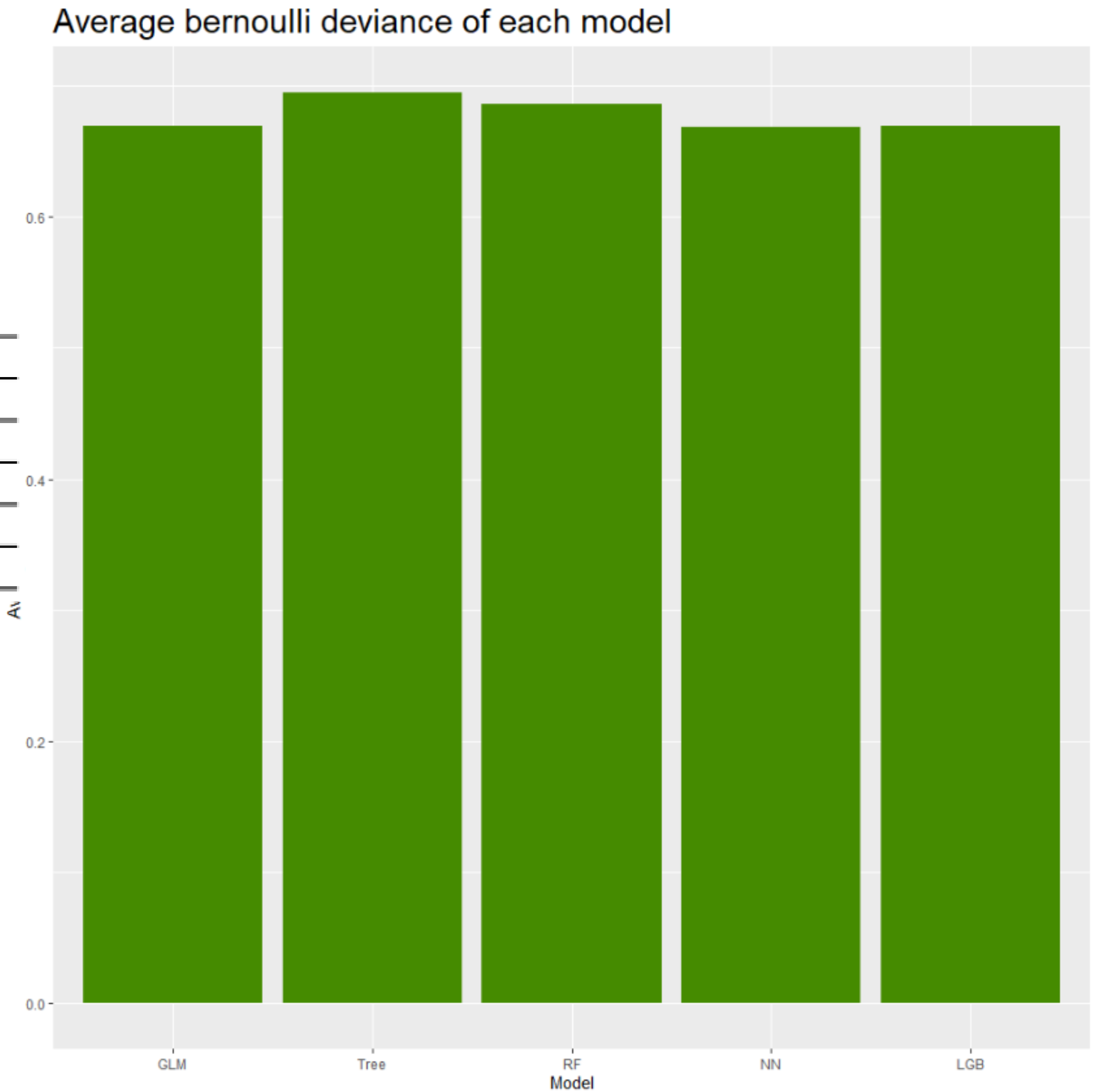
Mahdis Mousavi in Unsplash

Data

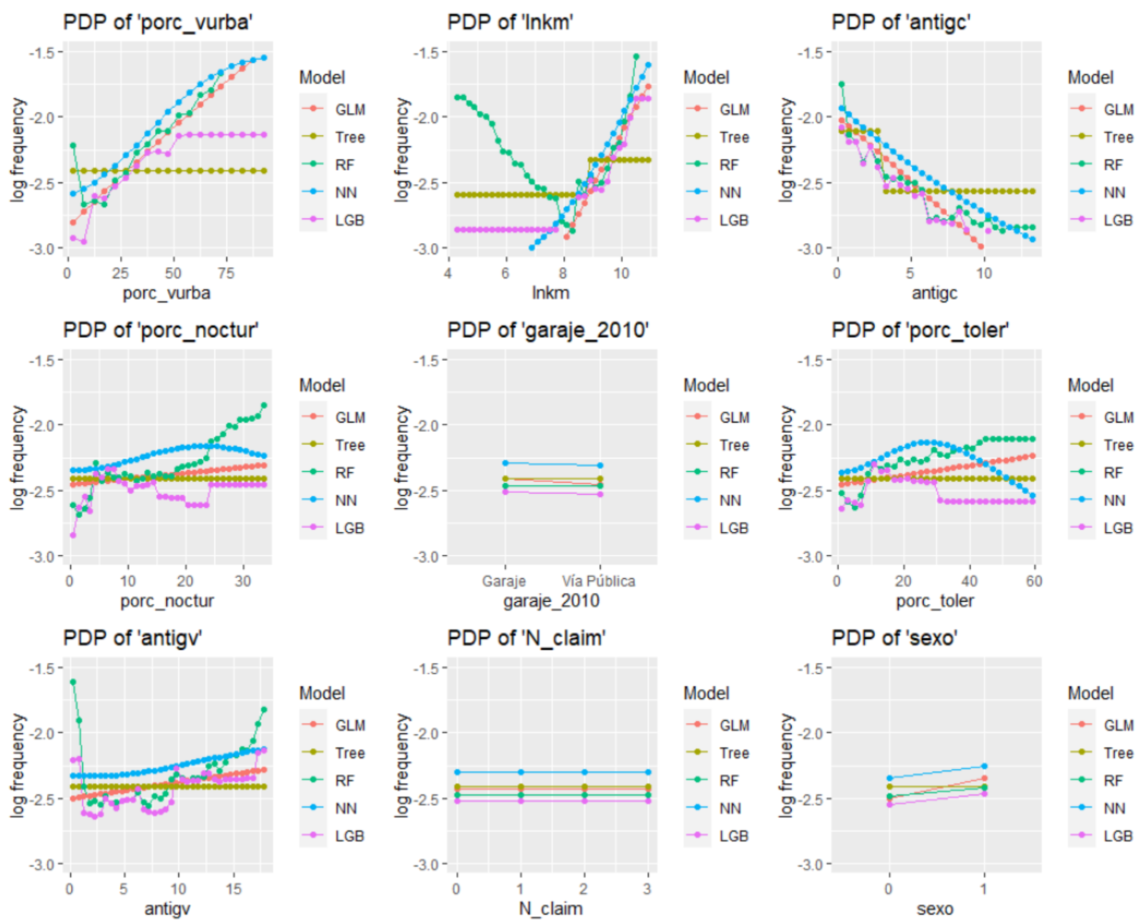
- Anonymous data were provided by a Spanish insurer that commercializes pay-as-you drive-insurance since 2009.
- Specifically, our data contain **9,614** drivers observed in Spain in 2010. One year aggregated information.
- A total of **926** claims at fault were observed.

Model comparison

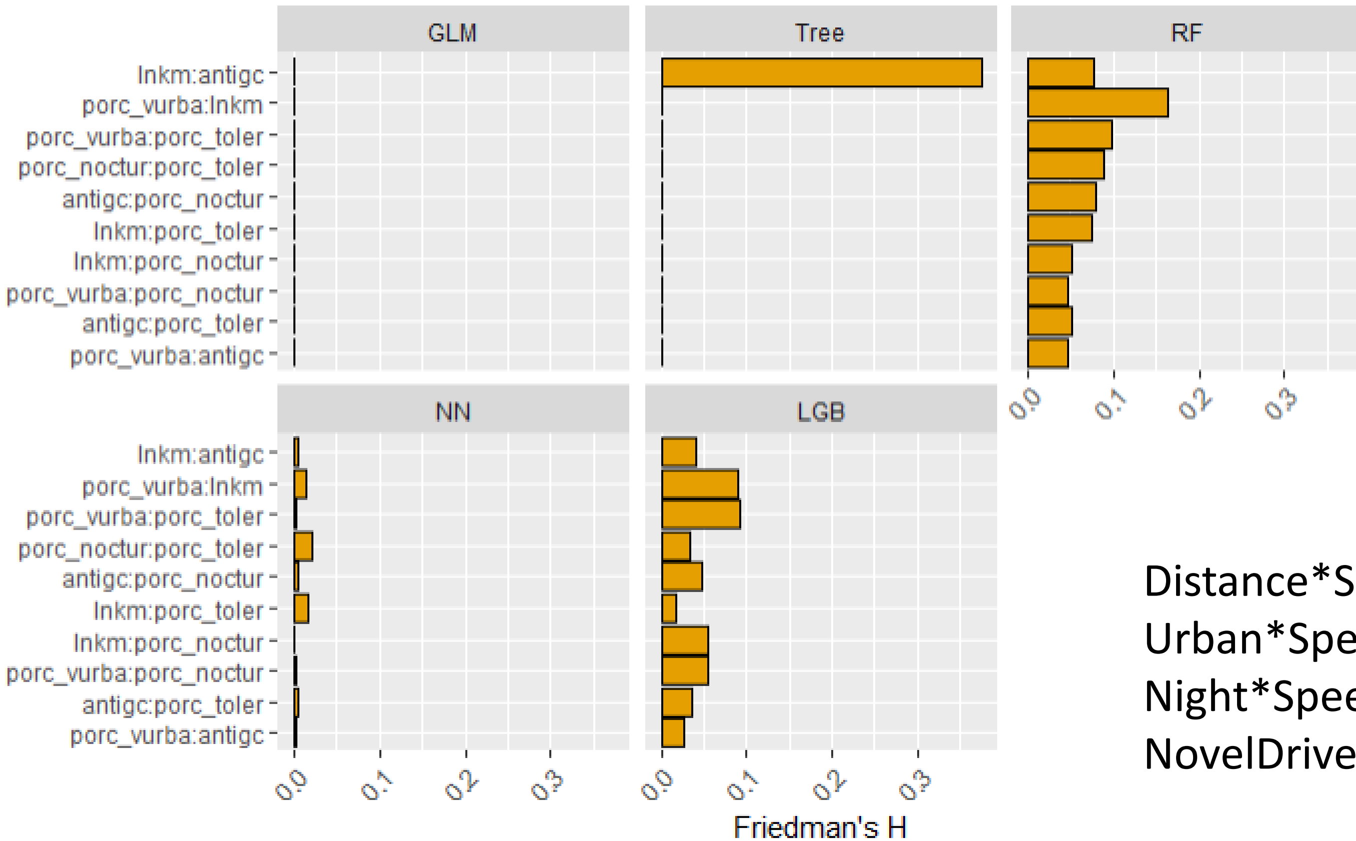
Model	Average Deviance Bernoulli
GLM	0.6696475
Tree	0.6949028
RF	0.6793481
NN	0.6691154
LGB	0.6673354



Partial Dependence Plot (PDP)

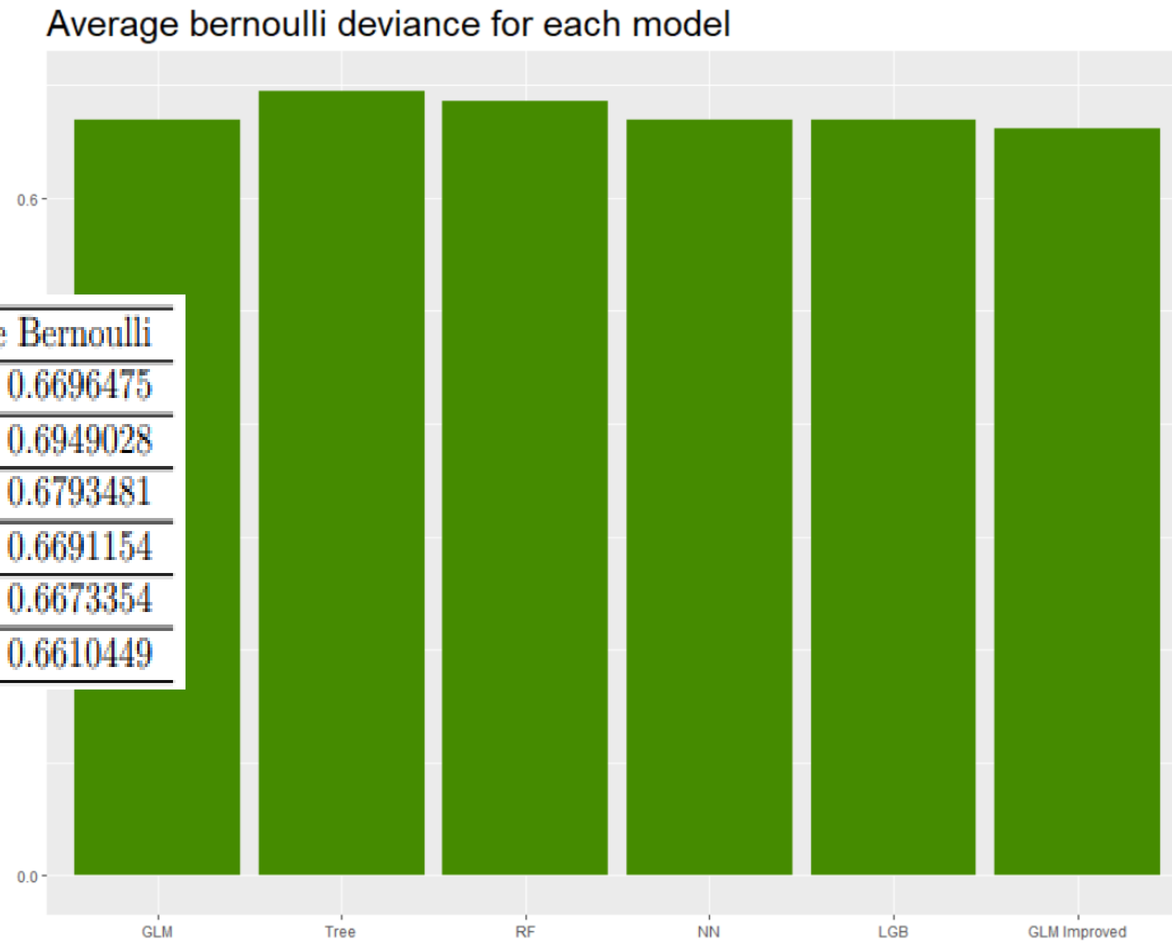


Identifying interactions



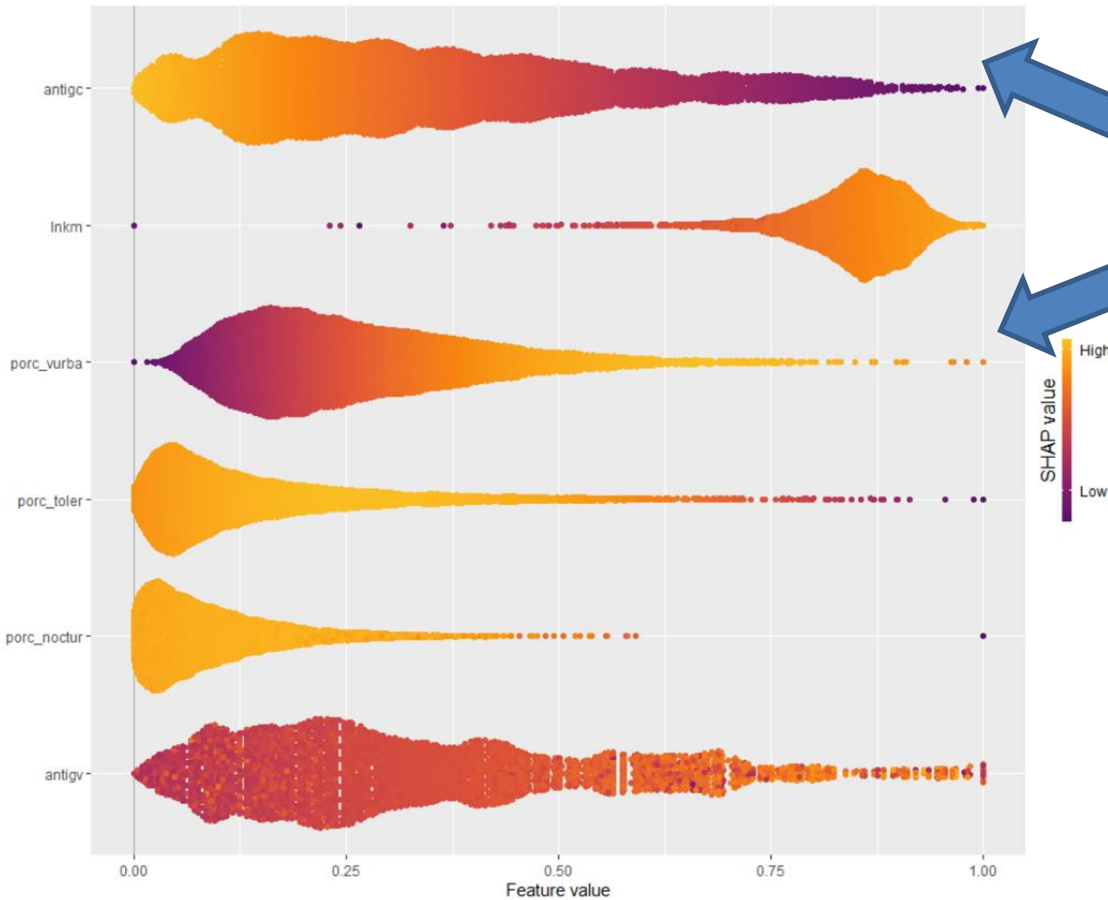
Distance*Speed
Urban*Speed
Night*Speed
NovelDriver*Night

Improving GLM with interactions



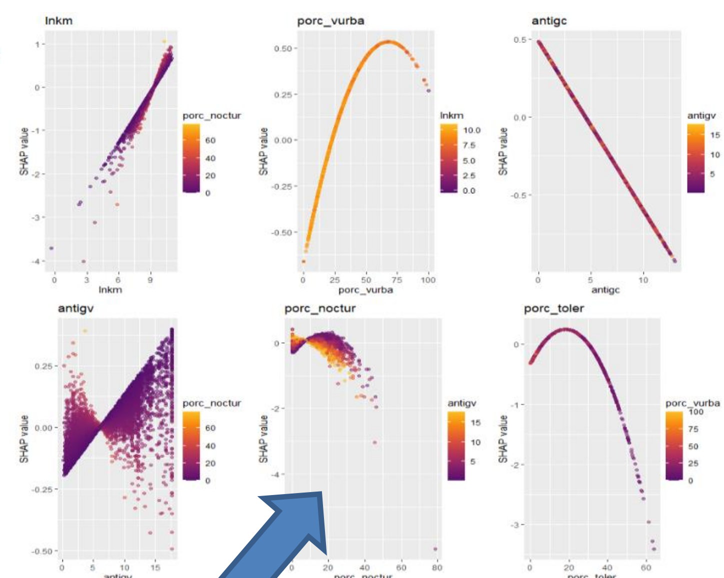
Model	Average Deviance Bernoulli
GLM	0.6696475
Tree	0.6949028
RF	0.6793481
NN	0.6691154
LGB	0.6673354
GLM Improved	0.6610449

SHAP (**SH**apley **A**dditive ex**P**lanation)



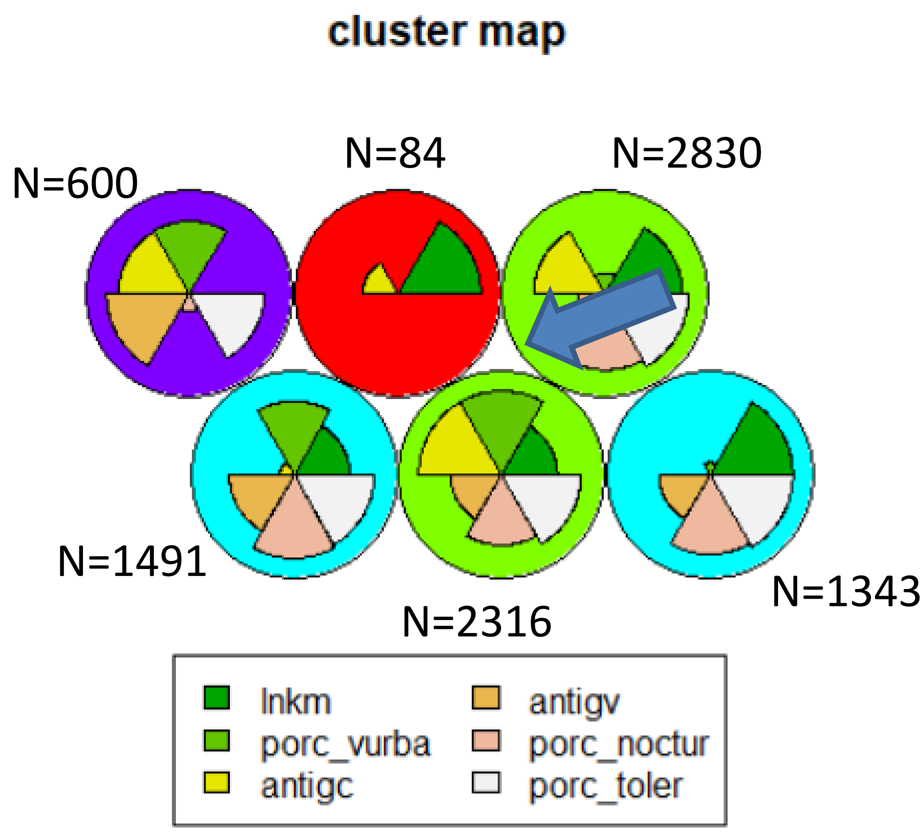
Novel driver vs senior driver

Non-urban vs urban drivers



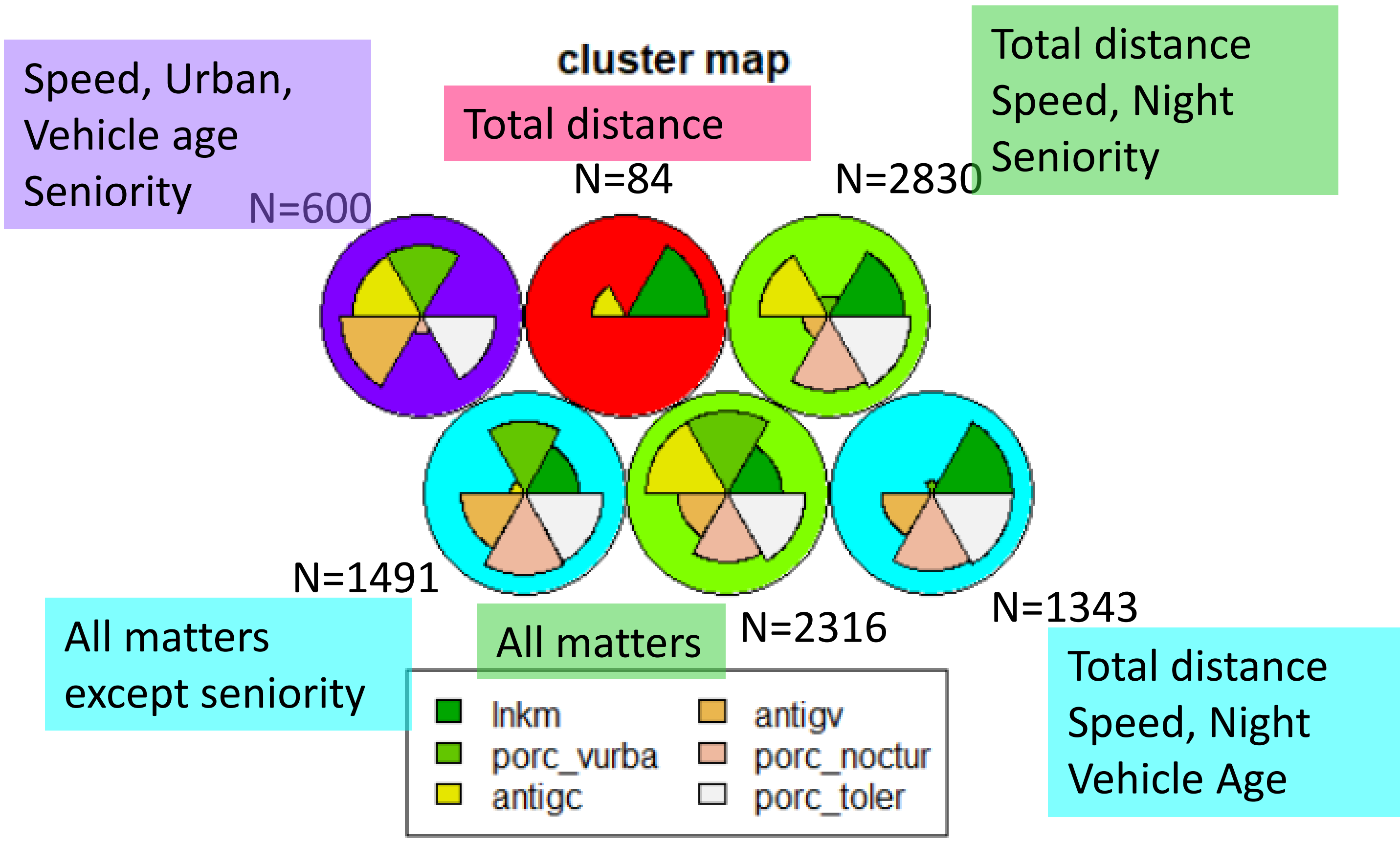
Shap value of vehicle age connected with night driving

Kohonen map from SHAP



950 drivers in smaller clusters

Kohonen map from SHAP



950 drivers in smaller clusters

Case study III: Take aways

- The GLM can be improved using ML methods
- Not all drivers are influenced by the same factors equally
- We can identify drivers that are more/less affected by telematics information than the rest

→ Identify policyholders that would benefit from different Pay-How You-Drive (PHYD) schemes.

CASE STUDY IV

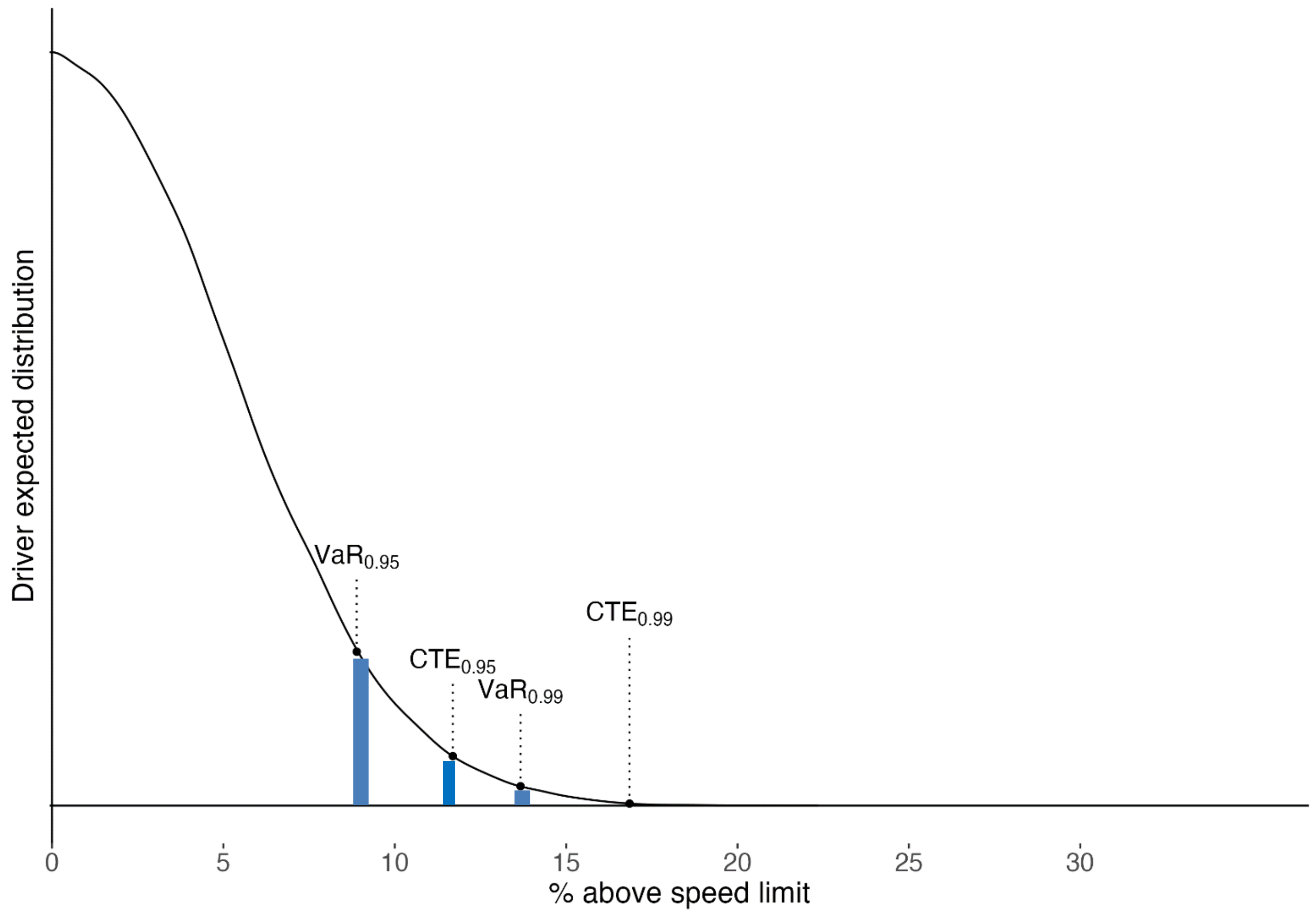
- Pricing with near-misses
- Weekly data
- Interpretable ML
- Scoring risky drivers

Hannes Egler in Unsplash

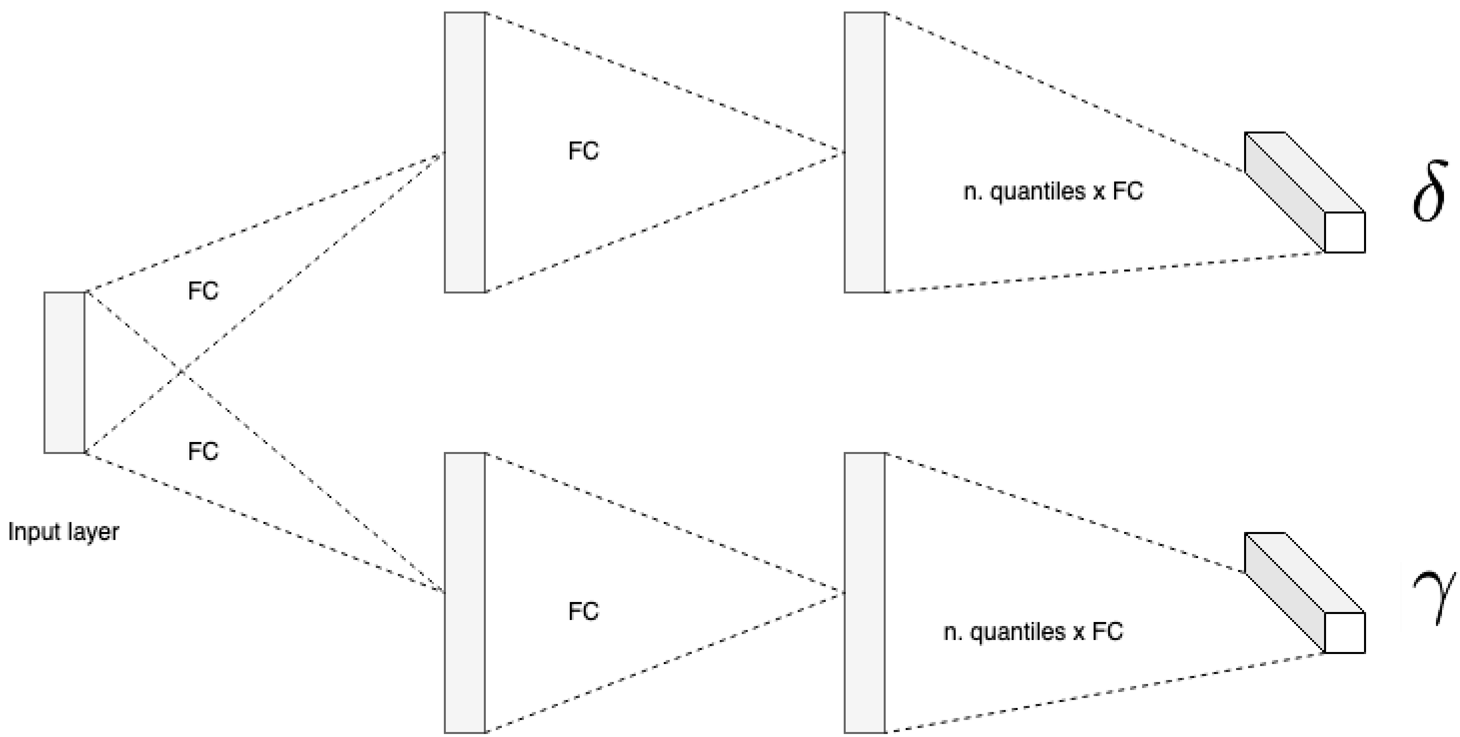
Data

- Anonymous data were provided by a Spanish insurer that commercializes pay-as-you drive-insurance since 2009.
- Specifically, our data contain **9,614** drivers observed in Spain in 2010.
- A total of **926** claims at fault were observed.

Quantile regression

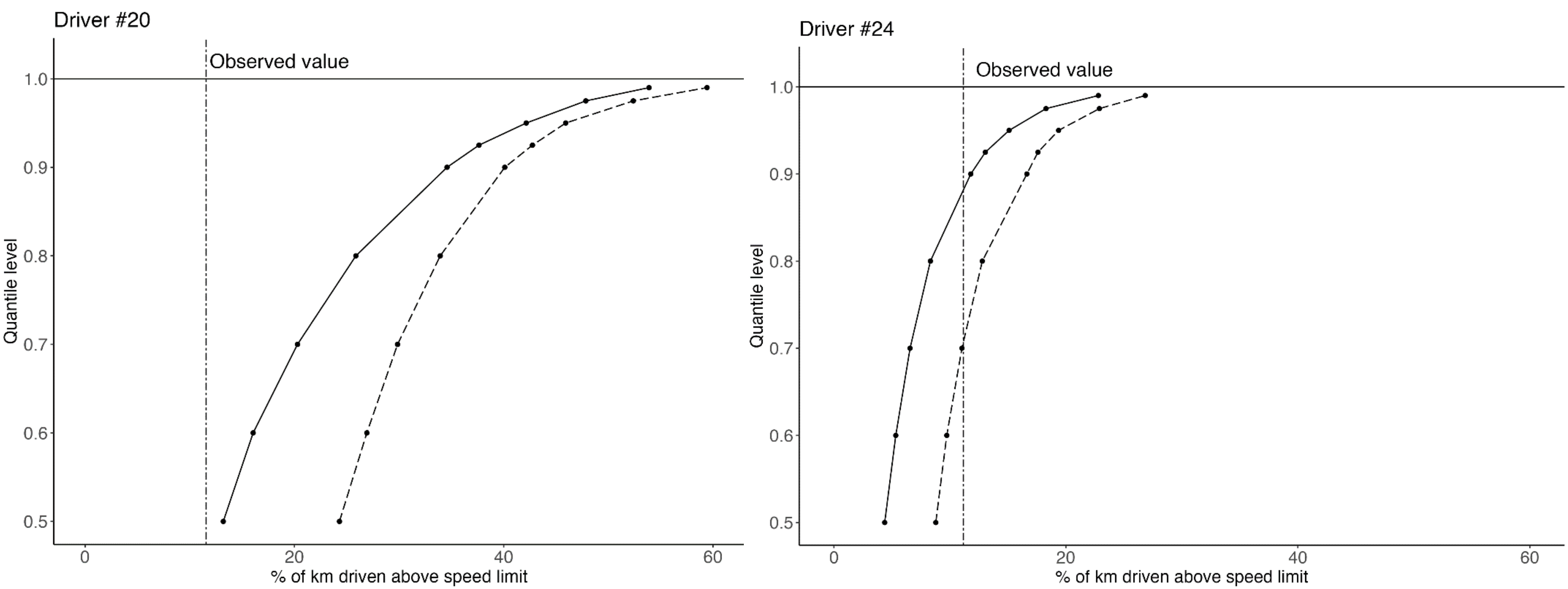


Neural network architecture aimed at quantile regression and CTE regression



Scoring drivers: conditional distribtion

#Driver	%km above speed limit	Age	Gender	Mileage (km)	% Urban driving	% Night driving
20	**11.57%**	25	Male	32,539	11.6%	4.8%
24	**11.16%**	24	Female	8,498	46.8%	12.7%



Case study IV: Take aways

- The new NN approach produces conditional quantiles regression results
- The estimated cumulative distribution is non-decreasing
- We can score drivers by locating their quantile level in the cumulative distribution function or in the CTE curve

→ Identify policyholders that require higher risk loadings in premium calculation.

Contents

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

- Maselo, L., Castignani, G., Sheehan, B., Guillen, M and Murphy, F. (2023) “Using contextual data to predict risky driving events: A novel methodology from Explainable Artificial Intelligence,” **Accident Analysis and Prevention** (accepted).
- 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 ,Ll. 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., Cot´e, S. and Guillen, M. (2018) “Exposure as duration and distance in telematics motor insurance using generalized additive models”, **Risks**, 5(4), 54.

Risk center RESEARCH

UNIVERSITAT DE BARCELONA

MINISTERIO DE CIENCIA E INNOVACIÓN

Financiado por la Unión Europea NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA ESTATAL DE INVESTIGACIÓN

ICREA

mguillen@ub.edu

Pricing motor insurance with telematics data

Montserrat Guillen

Chaire PARI, Novembre 15, 2023