Chaire Co-operators en
analyse des risques actuariels

How much telematics information do insurers need for claim classification?

Casualty Actuaries of Greater New York Spring 2021 Meeting

Francis Duval

May 12th 2021

Université du Québec à Montréal

Overview

Research question

When has an insurer collected enough information about an insured's driving habits ?

General idea

- Development of a claim classification model using **telematics** data.
- Development of a method based on **claim classification** to determine when telematics information becomes redundant.

Motivations

- An insurer wishes to keep a minimum of telematic information on its policyholders for reasons of :
  - Confidentiality
  - Data storage
- But still wants to take advantage of this information, for instance, to avoid adverse selection.

Trip data

Extract from the trip database

VIN	Trip ID	Starting time	Arrival time	Distance	Maximum speed
A	1	2016-04-09 15:23:55	2016-04-09 15:40:05	10.0	72
A	2	2016-04-09 17:49:33	2016-04-09 17:57:44	4.5	68
⋮	⋮	⋮	⋮	⋮	⋮
A	3312	2019-02-11 18:33:07	2019-02-11 18:54:10	9.6	65
B	1	2016-04-04 06:54:00	2016-04-04 07:11:37	14.0	112
B	2	2016-04-04 15:20:19	2016-04-04 15:34:38	13.5	124
⋮	⋮	⋮	⋮	⋮	⋮
B	2505	2019-02-11 17:46:47	2019-02-11 18:19:22	39.0	130
C	1	2016-01-16 15:41:59	2016-01-16 15:51:35	3.3	65
⋮	⋮	⋮	⋮	⋮	⋮

- These are the only telematics data we have. All telematics features are derived from these **4 measurements**.

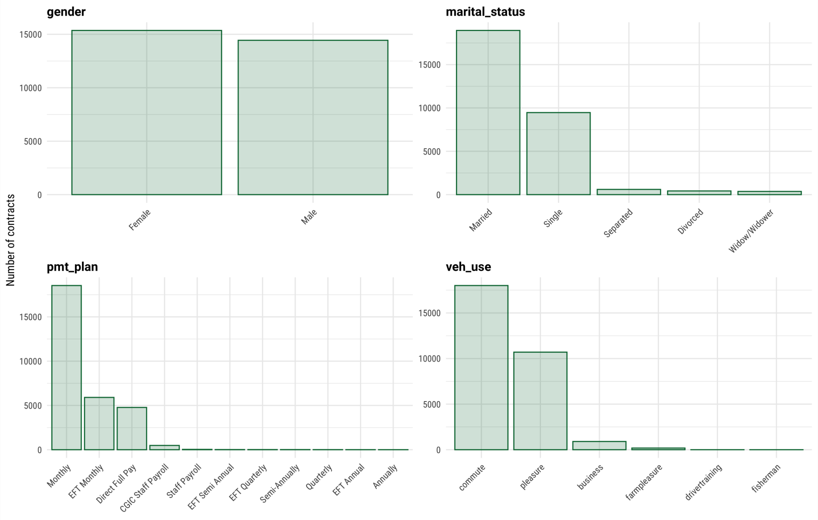
Contract data

Extract from the contract database

VIN	Contract start date	Contract end date	Classic covariate #1	...	Claim(s) indicator
A	2015-01-09	2016-01-09	F	...	0
A	2016-01-09	2017-01-09	F	...	1
A	2017-01-09	2018-01-09	F	...	0
B	2015-12-14	2016-12-14	M	...	0
B	2016-12-14	2017-12-14	M	...	0
C	2015-04-26	2016-04-26	F	...	1
C	2016-04-26	2017-04-26	F	...	0
C	2017-04-26	2018-04-26	F	...	0
⋮	⋮	⋮	⋮	⋮	⋮

- Linking of the 2 datasets on the basis of the VIN and the start/end dates of the contract.
- Expansion of the contract database with **14 telematics features** calculated using the trip dataset.

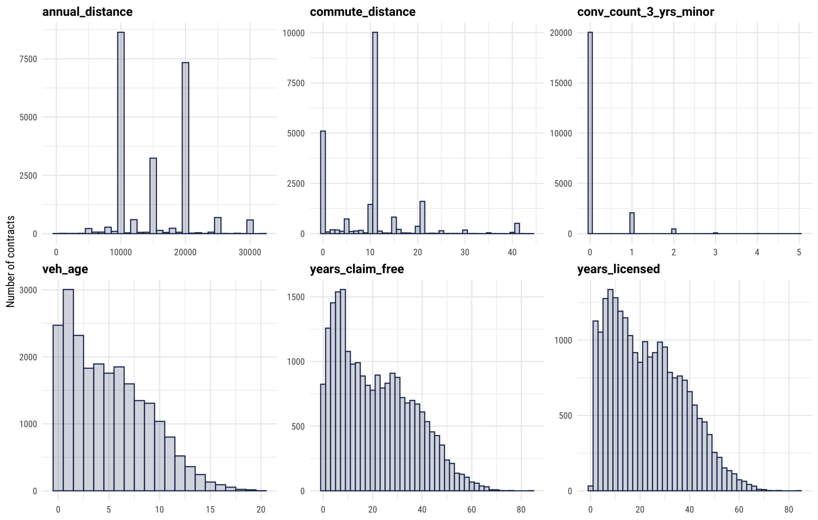
Classic features – Categorical



Preprocessing :

Lump rare categories ⟶ target encode ⟶ normalize ⟶ Yeo-Johnson transform

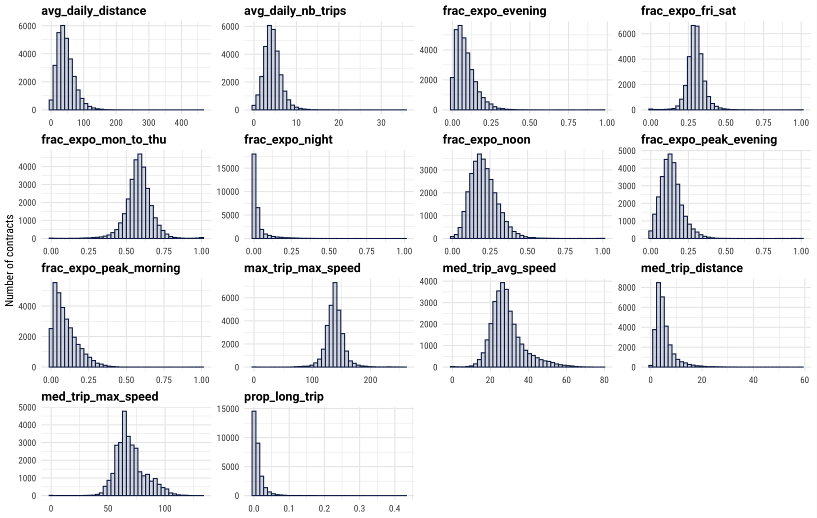
Classic features – Numeric



Preprocessing :

Normalize ⟶ Yeo-Johnson transform

Telematics features



Preprocessing :

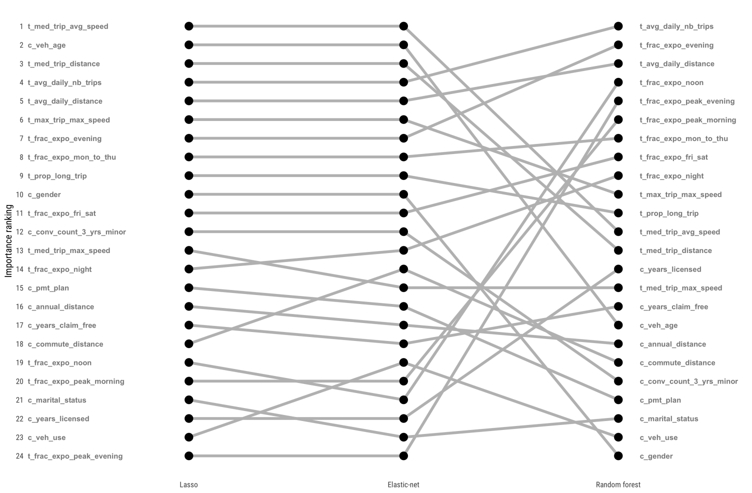
Normalize ⟶ Yeo-Johnson transform

We consider 3 classification algorithms :

- Lasso logistic regression
- Elastic-net logistic regression
- Random forest

	Optimal hyperparameters					
Models	λ	α	p^*	n^*	**AUC (5-fold cross-validation)**	**AUC (testing set)**
Lasso	2.31×10^{-4}	–	–	–	$0.6373^{(0.0052)}$	0.6189
Elastic-net	2.98×10^{-3}	0	–	–	$0.6377^{(0.0049)}$	0.6176
Random forest	–	–	1	39	$0.6004^{(0.0064)}$	0.5889
Lasso (with interactions)	1.18×10^{-3}	–	–	–	$0.6350^{(0.0050)}$	0.6214
Elastic-net (with interactions)	1.52×10^{-2}	0	–	–	$0.6359^{(0.0046)}$	0.6198

Feature importance



- Top 10 features are almost all telematics.
- Some of the most important features are **t_avg_daily_nb_trips**, **t_avg_daily_distance**, **t_med_trip_avg_speed**, **t_max_trip_max_speed**, **t_frac_expo_evening** and **t_frac_expo_mon_to_thu** and **c_veh_age**.

A glimpse at lasso logistic regression

Loss function

$$L(\boldsymbol{\beta},\mathbf{y}) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\} + \lambda\sum_{j=1}^{p}|\beta_j|, \quad \text{where} \quad p_i = \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i^\top \boldsymbol{\beta}}}$$

Estimation

- We find the $\boldsymbol{\beta}$ coefficients that minimize the loss function, which is equivalent to minimizing the negative of the log-likelihood with a constraint on the sum of the absolute values of the coefficients :

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\arg\min}\left\{-\frac{1}{n}\sum_{i=1}^{n} y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

Prediction

- Same prediction formula as a non-penalized logistic regression, but using lasso coefficients $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$:

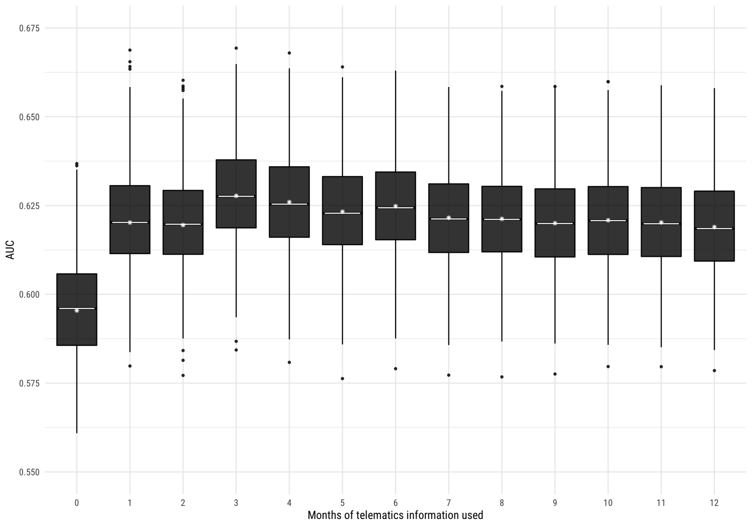
$$\widehat{y}_i = \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{\text{lasso}}}}$$

Methodology

1. Create **k versions of the telematics features** using varying amounts of trip summaries for each vehicle.
2. Create **k classification datasets** derived from these k versions of telematics features and the **classic features** plus a classification dataset with only **classic features**. Split each of them into **training** and **testing** sets.
3. **Tune and train** a lasso classification model on each of the $k+1$ **training** datasets.
4. **Assess the performance** of the $k+1$ models on their respective **testing** dataset.

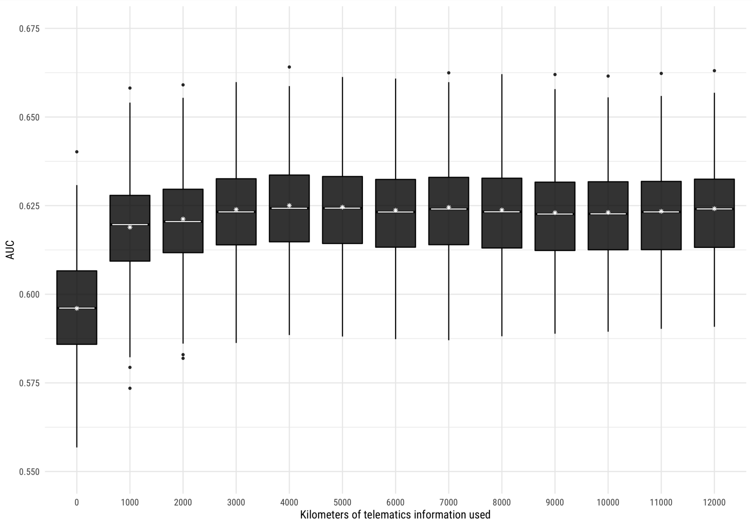
- We choose to create **12 versions** of the telematics features, each using **one month more data** than the previous version.
- We therefore have **13 classification datasets**.
- We assess the performance using the **AUC**. In order to obtain a **distribution** of this performance metric, we use **non-parametric bootstrapping**.

Results – Time leaps



- The AUC has improved substantially with the 4-measure trip summaries !
- Telematics information becomes redundant after about **3 months**.

Results – Distance leaps



- Telematics information becomes redundant after about **4,000 km**.

Conclusions

Summary

- We have developed a **claim classification model** using **telematics** data in the form of **trip summaries**.
- Based on this claim classification model, we have designed a **method useful to determine when information on the insured's driving becomes redundant**.
- With the data we have at hand, we found out that telematics information no longer improves classification performance after about **3 months** or **4,000** km of trip summaries.

Future considerations

- Do we come to the same conclusions if we use, for instance, comprehensive coverage claims (theft, hail, etc.) ?
- Generalize the approach for count regression.