Pricing using the address:

High dimensional data set with heterogeneous quality

To avoid adverse selection, insurance companies seek to have the most precise premium. In household insurance pricing, the models components stems from questionnaire with in average 20 questions. The more questions or the more complex are the questions, the more precise is the premium but the less quotation are done. Using the address to access computer vision and external data - type of roof, presence of windows on roof, house value, ... - could household insurance be priced without questionnaire?

1. Buildings geolocalisation

Digital process: Quick Underwriting process

More variables and more complexe information available. Quicker underwriting process meaning more quotation finished.

2. Data selection per risk

To limit the impact of geolocalisation errors, data selection constraints are optimised by coverage/risks.

3. Variables selection per risk

Methodology

Phase 1: Avoid redundancty through PCA and correlation matrix.
Phase 2: Filter selection using Random forest and GLM Lasso.
Phase 3 & 4: Pricing using GLM models (Log Gamma and Poisson) due to operational constraints.

4. Comparison between Traditional models and models using building’s geolocalisation

Traditional models: Models based on questionnaire variables

Quick underwriting models: Models based only on external data given by the building’s geolocalisation

Performance models: Models based on questionnaire variables and on external data

Results

- Adding new variables thanks to geolocalised building, the Performance model is significantly improving.
- Without questionnaire, Pricing models are as performing as the Traditional models.