KU LEUVEN

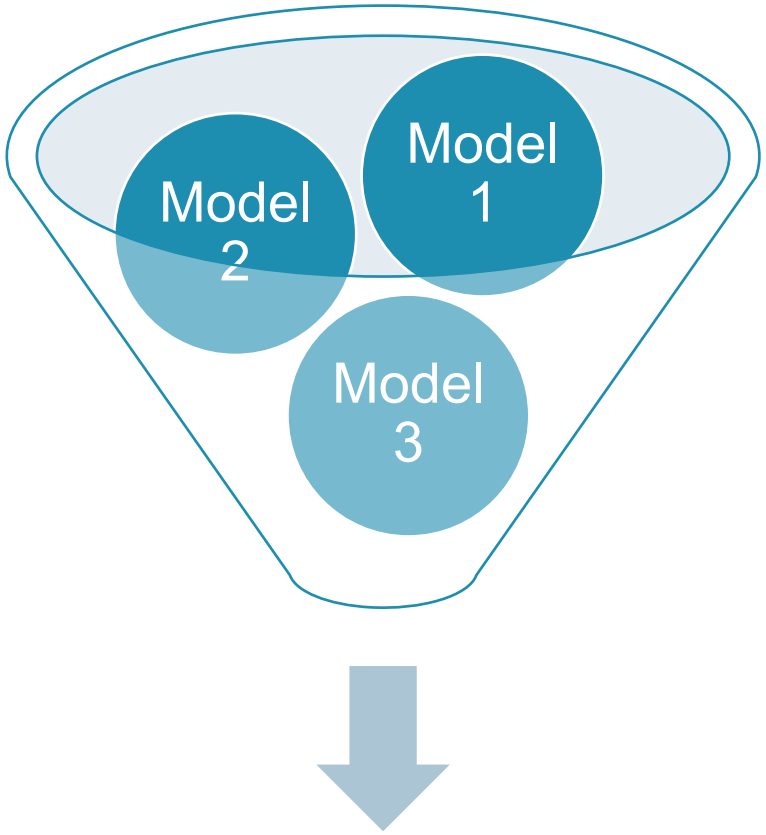
Model boosting and stacking for insurance pricing

Master thesis of Ine Fransen

Master of Actuarial and Financial Engineering

Prof. Dr. K. Antonio, Ir. R. Henckaerts

Ensemble approaches combine different models



Ensemble Model

(e.g. Model Boosting, Stacking)

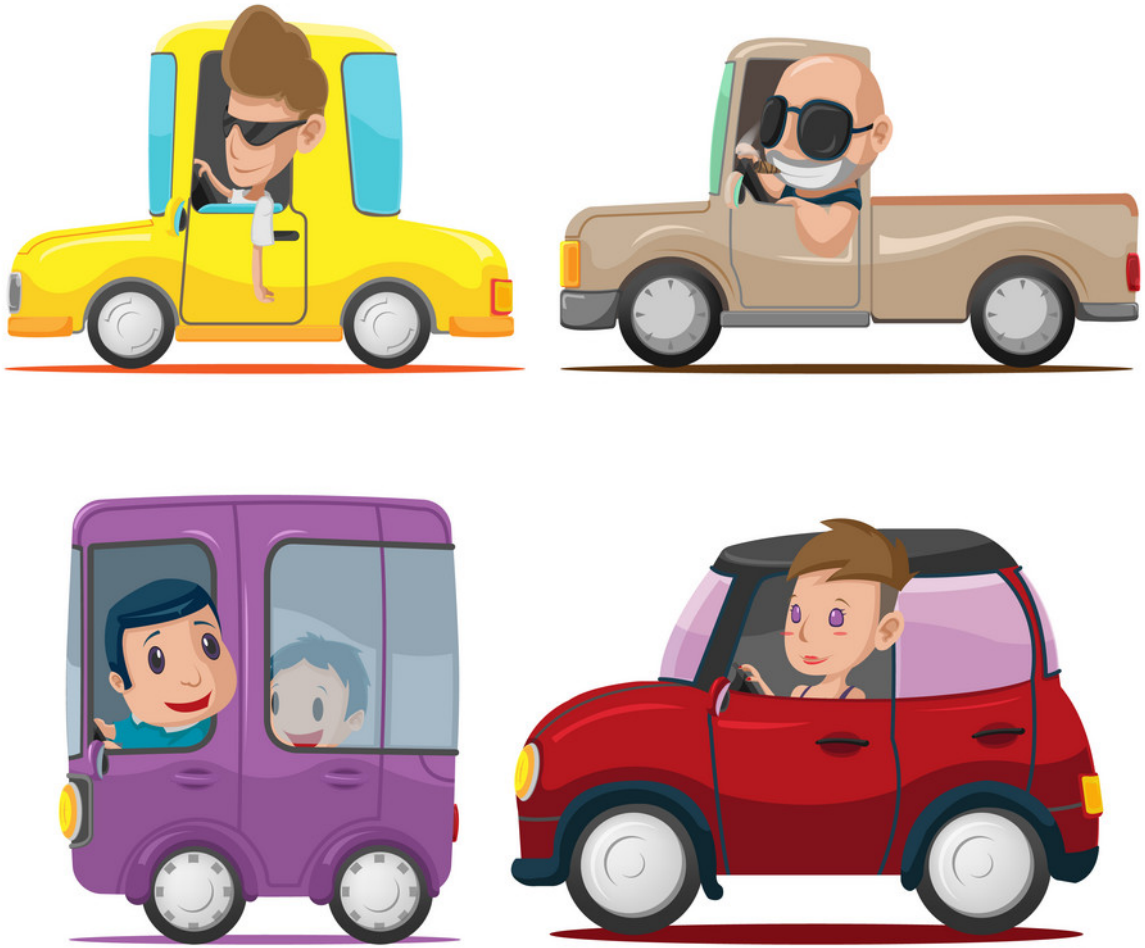
Goals of the thesis

- Present the concept of **model boosting**
- Present the concept of **stacking**

Apply to car insurance pricing

- Compare model performance and interpretability

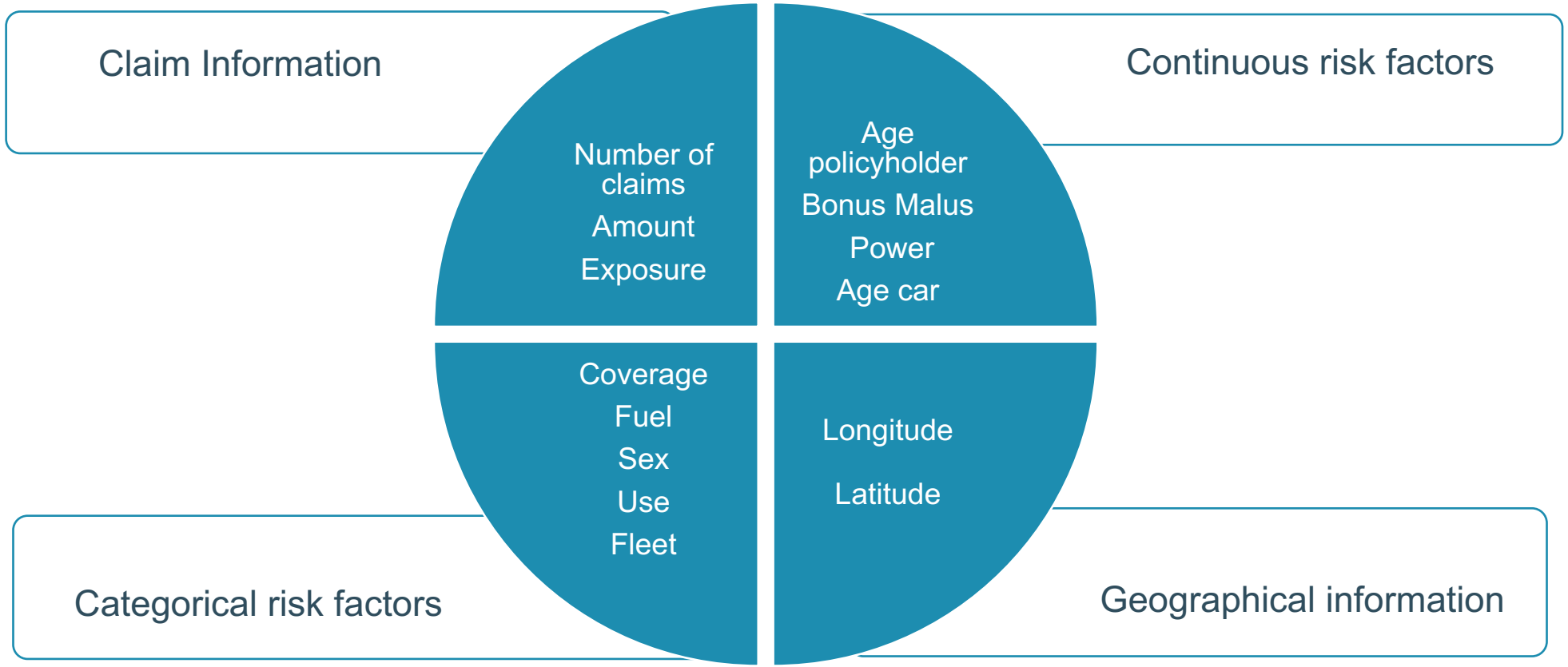
Portfolio in non-life insurance



https://www.vectorstock.com/royalty-free-vector/cars-driver-cartoon-collection-set-vector-14066666

Construction of a pricing model

Motor Third Party Liability (MTPL) data 1997



Calculation of the premium

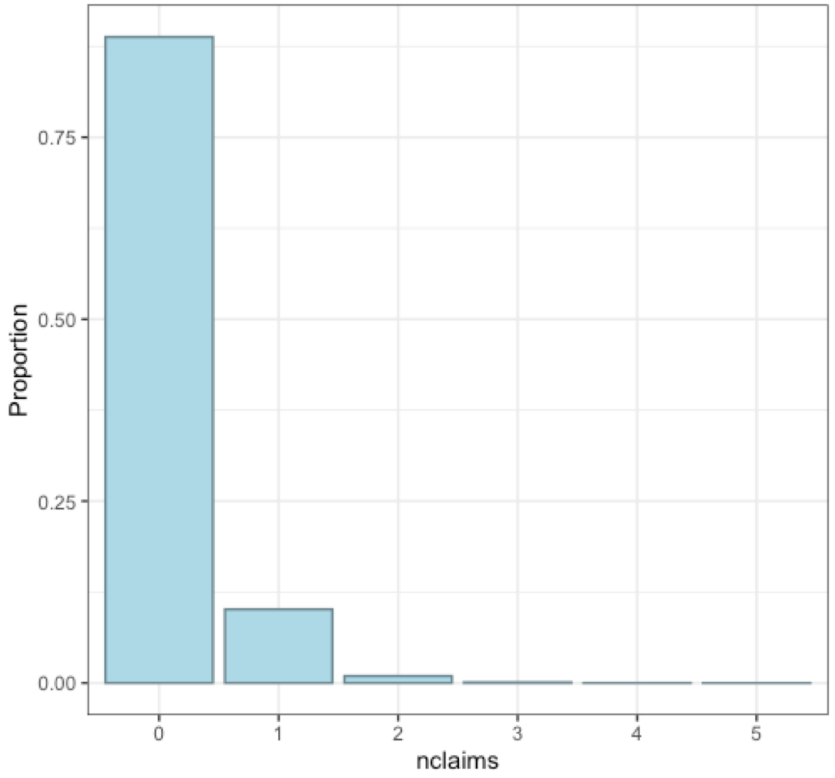
$$\pi_i = \mathbb{E}[F_i] \times \mathbb{E}[S_i]$$

Frequency

Severity

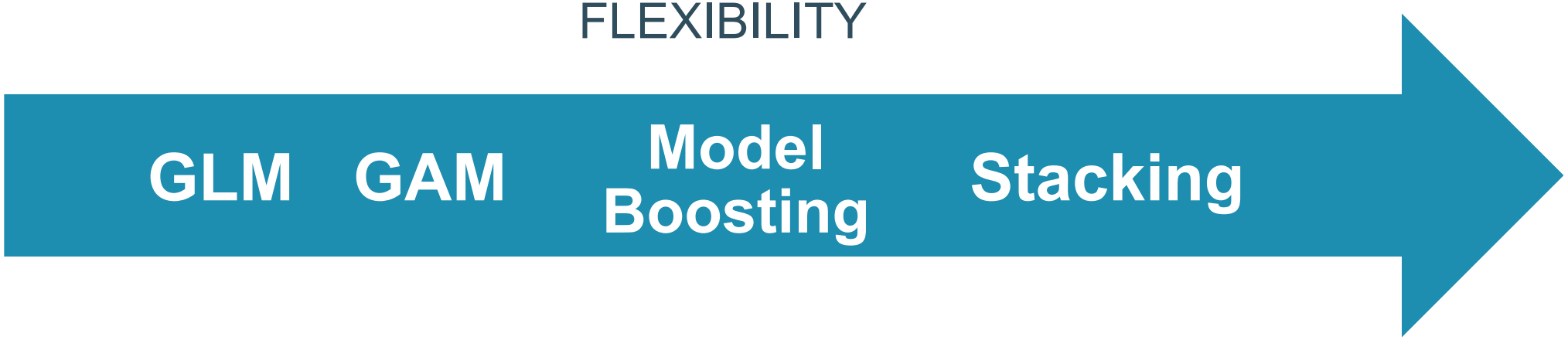
Challenges

- Selection of relevant variables
- Different types of risk factors
- Interpretability of the pricing model
- Distribution of the target variable



Demand for flexible yet explainable models

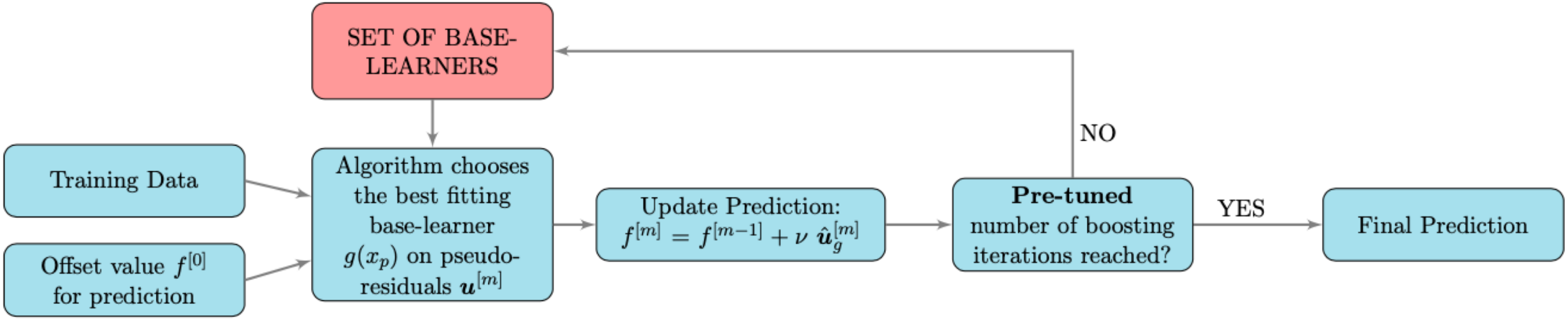
FLEXIBILITY



Explainable

Black Box

Model Boosting



Model Boosting

- Base-learners:
  - **Linear** effects: e.g. $\beta\, sex$
  - **Smooth** effects: e.g. $\sum_{j=1}^{t} B_j(age, q)$
  - **Tree-based** effects: e.g. $\sum_{j=1}^{J_1} \hat{y}_{R_j} \mathbb{I}(age \in R_j)$
- Automatic variable selection!

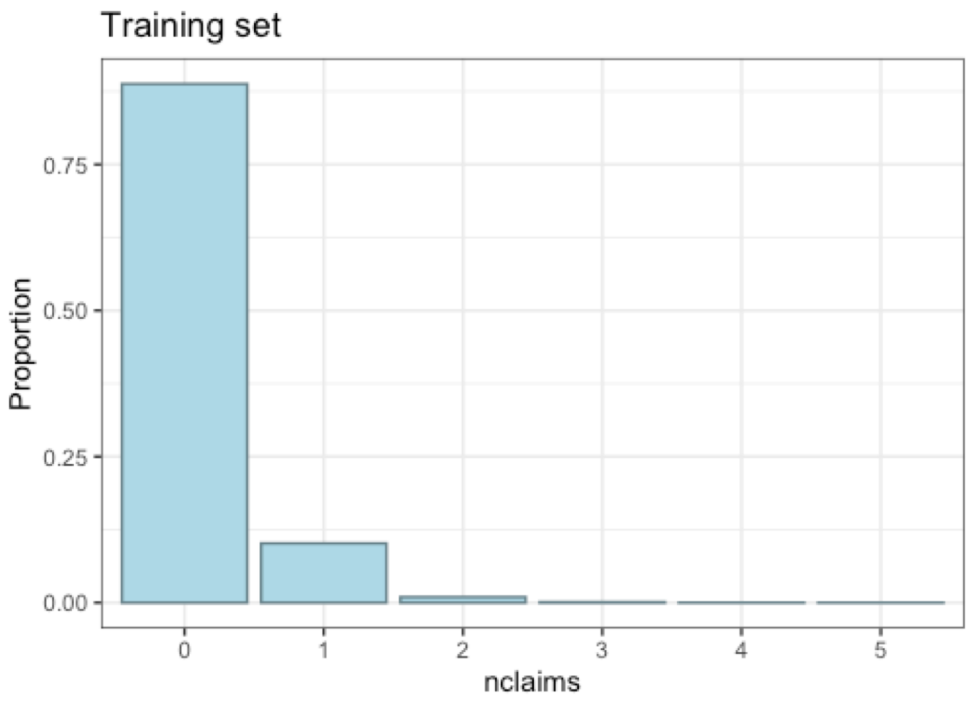
Three boosting models for claim frequency

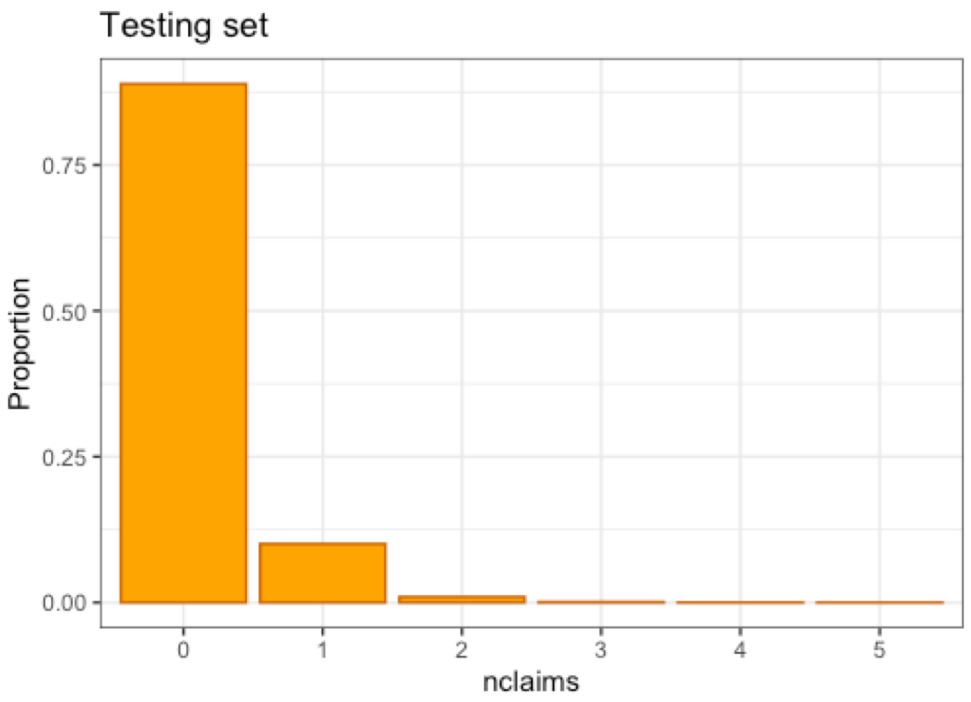
- Smooth model
- Tree model
- Hybrid model

Smooth base-learners: continuous risk factors & spatial effect

Tree-based base-learners: categorical risk factors & interactions

Training and test set

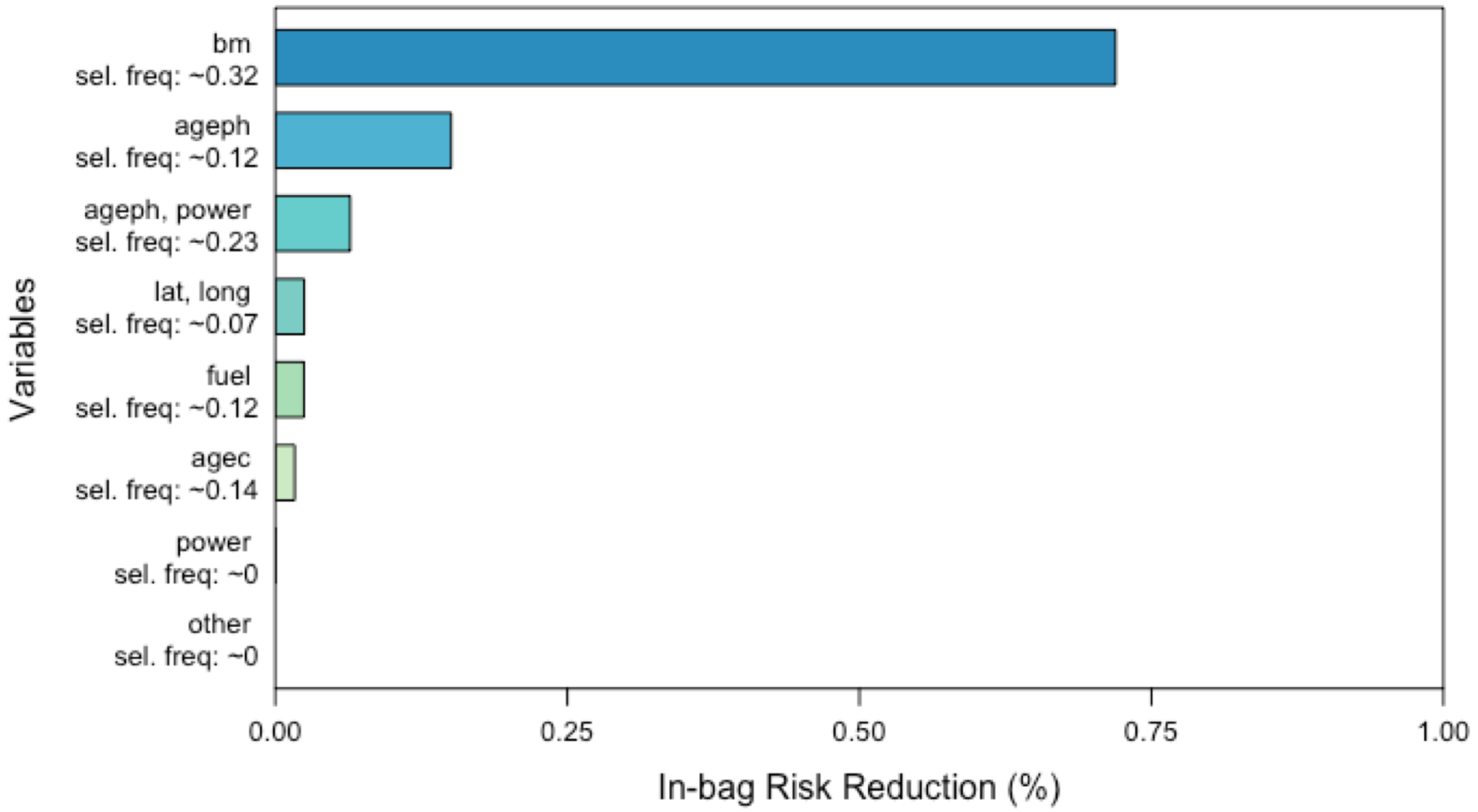




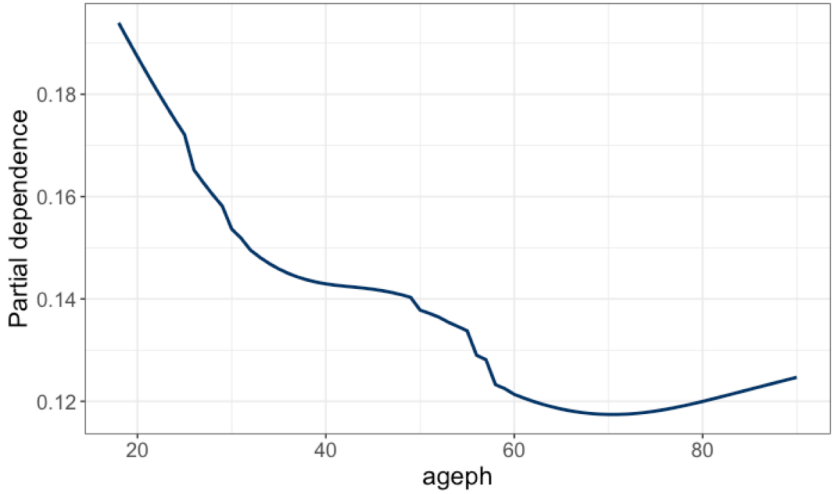
Comparison with GLM

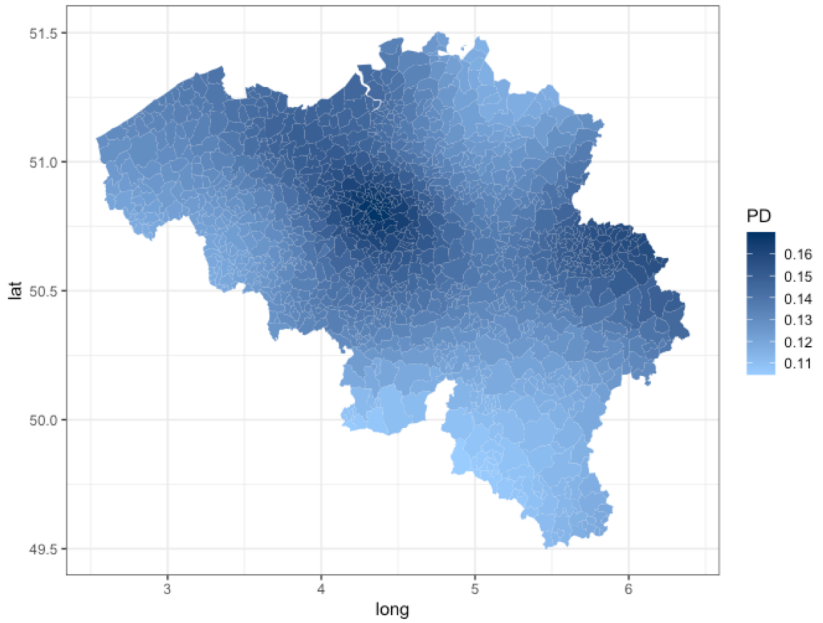
- Boosting can be more accurate
- Boosting is computationally costly
- Interpretability of boosting can be increased using
  - variable importance
  - partial dependence plots

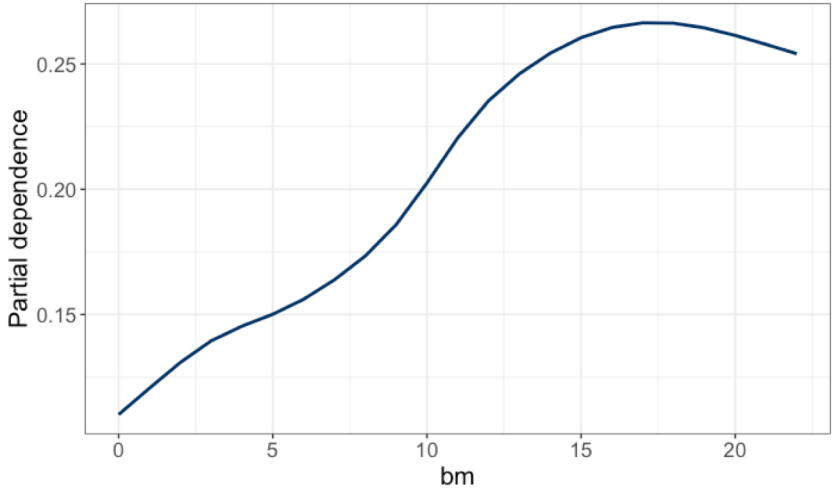
Variable importance

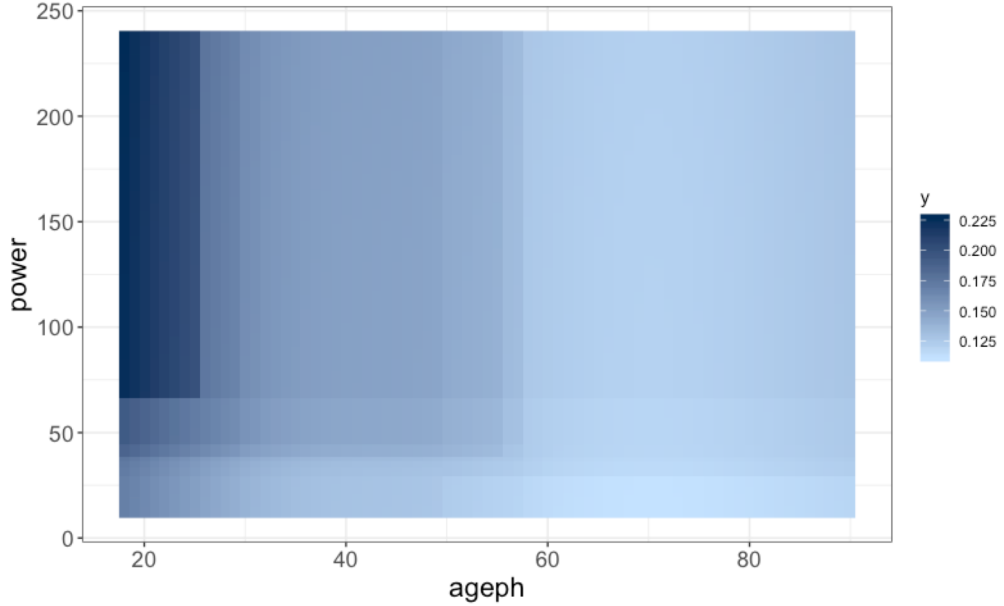


Partial dependence plots

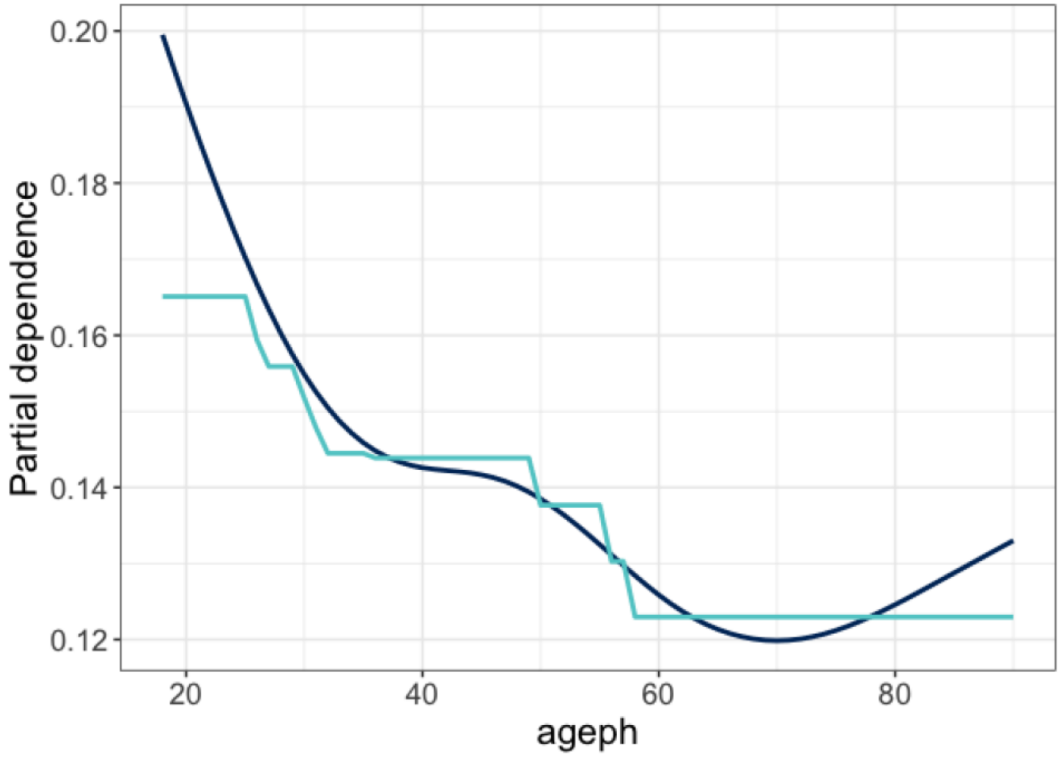


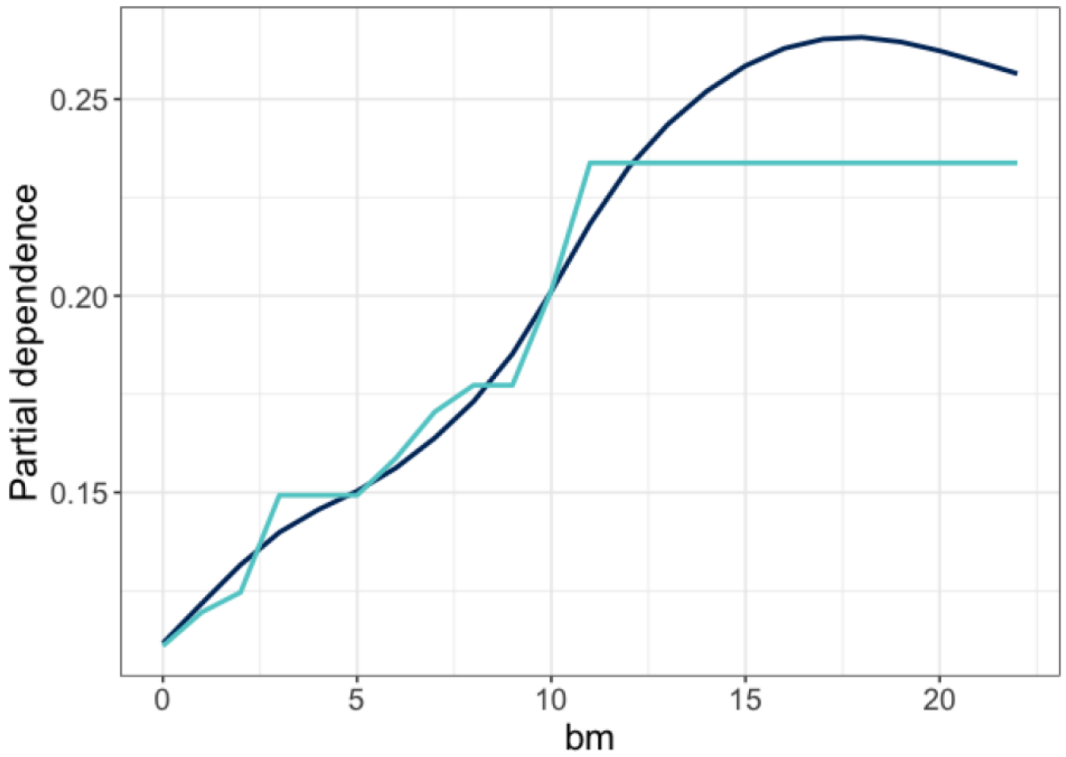




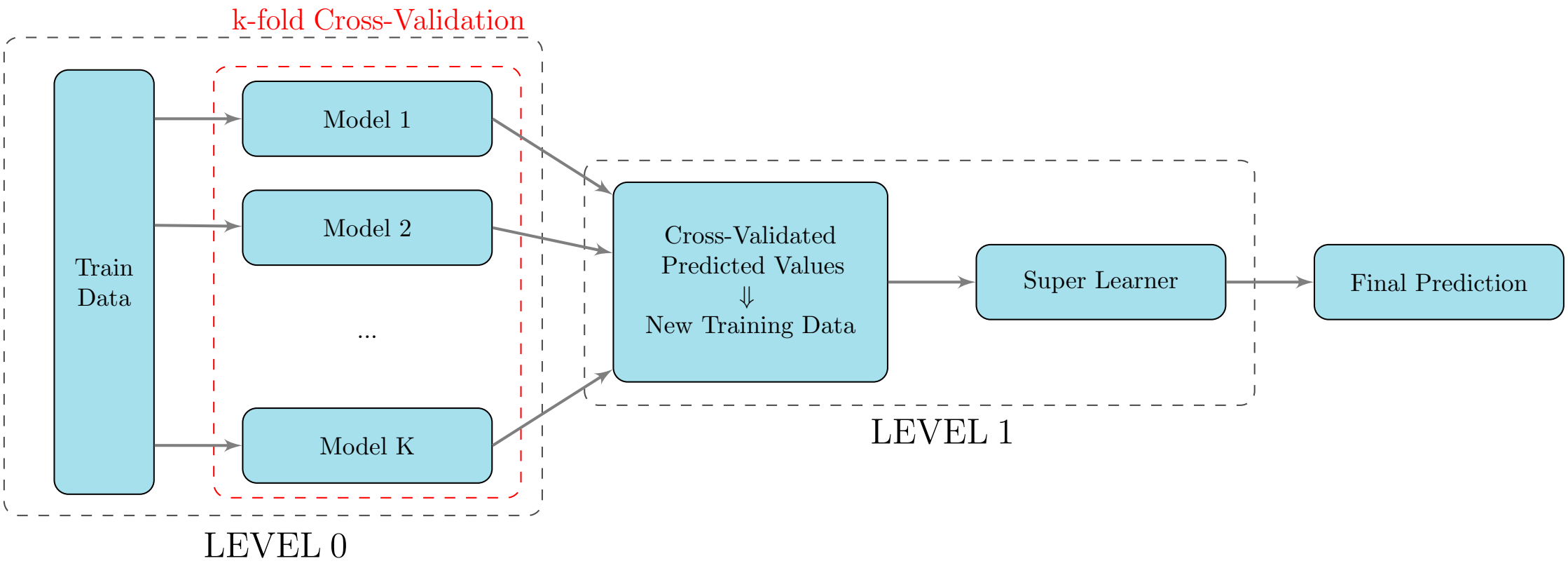


Approach for automatic binning





Stacking


k-fold Cross-Validation
Train Data
Model 1
Model 2
...
Model K
Cross-Validated Predicted Values
⇓
New Training Data
Super Learner
Final Prediction
LEVEL 1
LEVEL 0

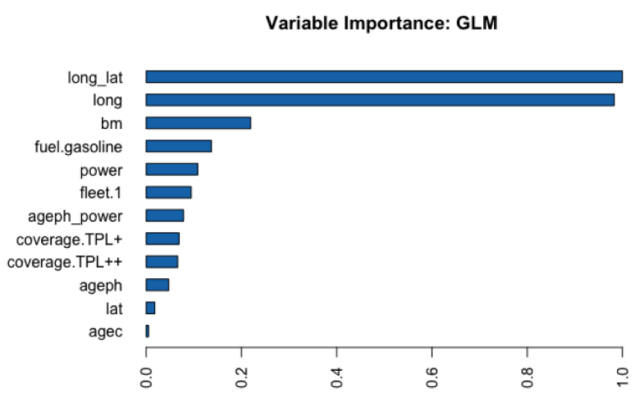
Four different models are stacked

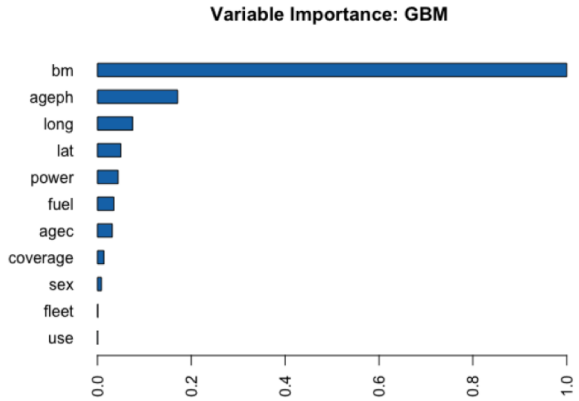
- GLM
- GBM
- XGBoost
- Random Forest

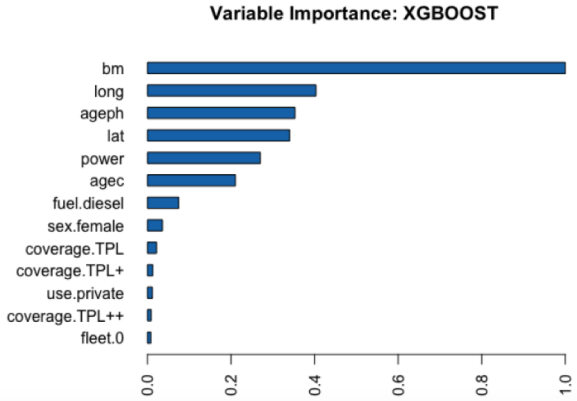
Comparison

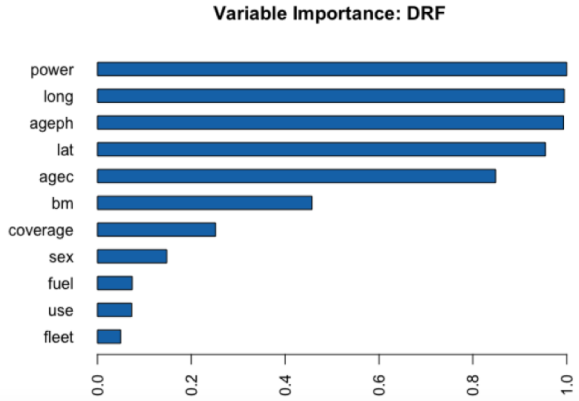
- Better accuracy than single models (GLM, Boosting models,...)
- Reduced interpretability
  - Variable importance
  - Partial dependence plots
- Increased computation time
- Overkill?

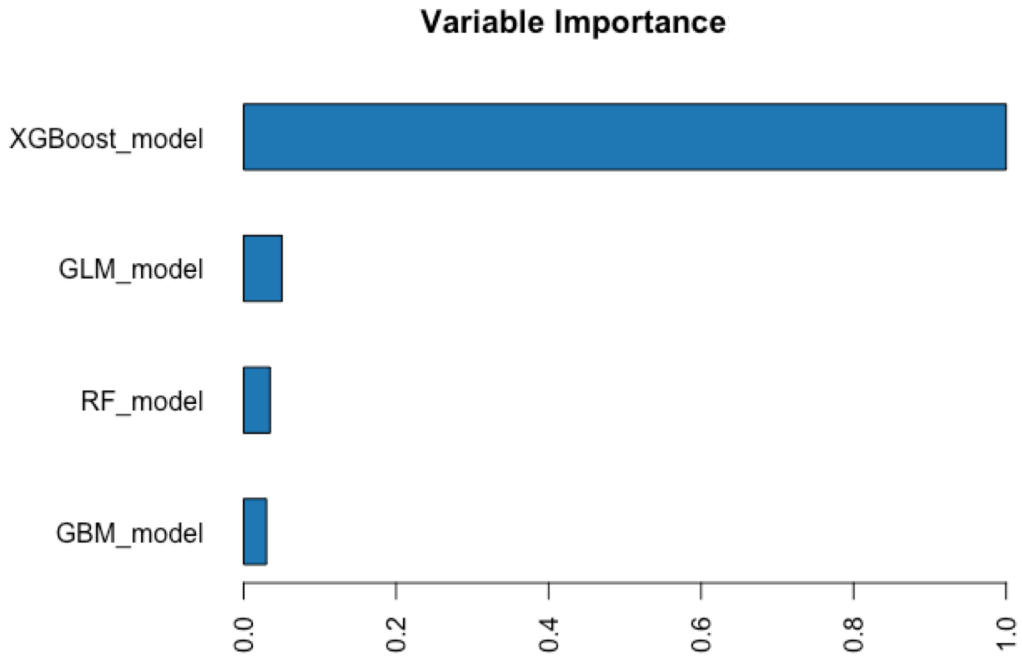
Variable importance



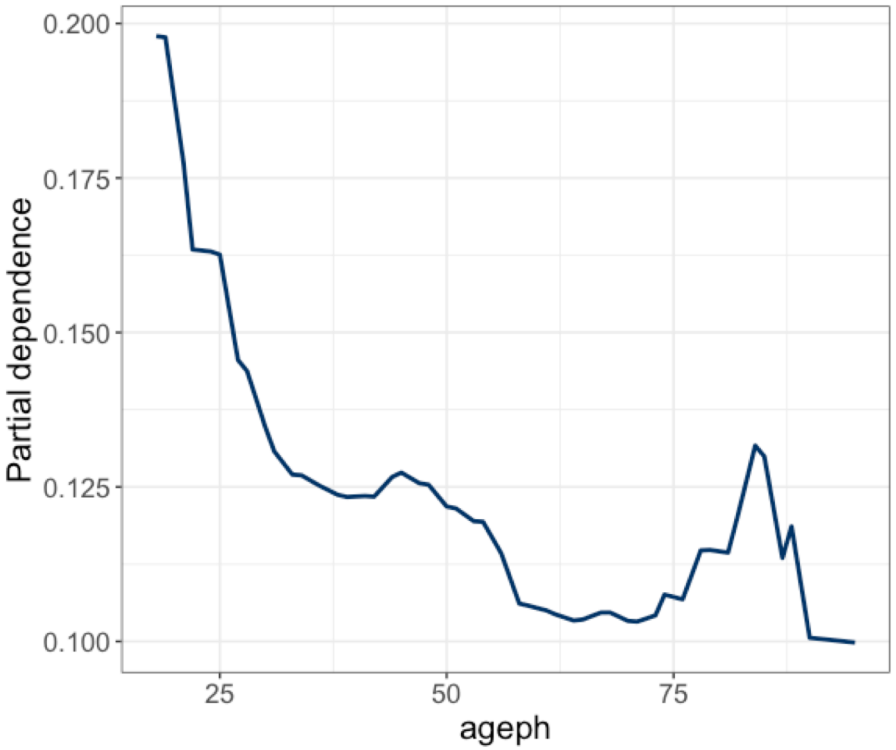


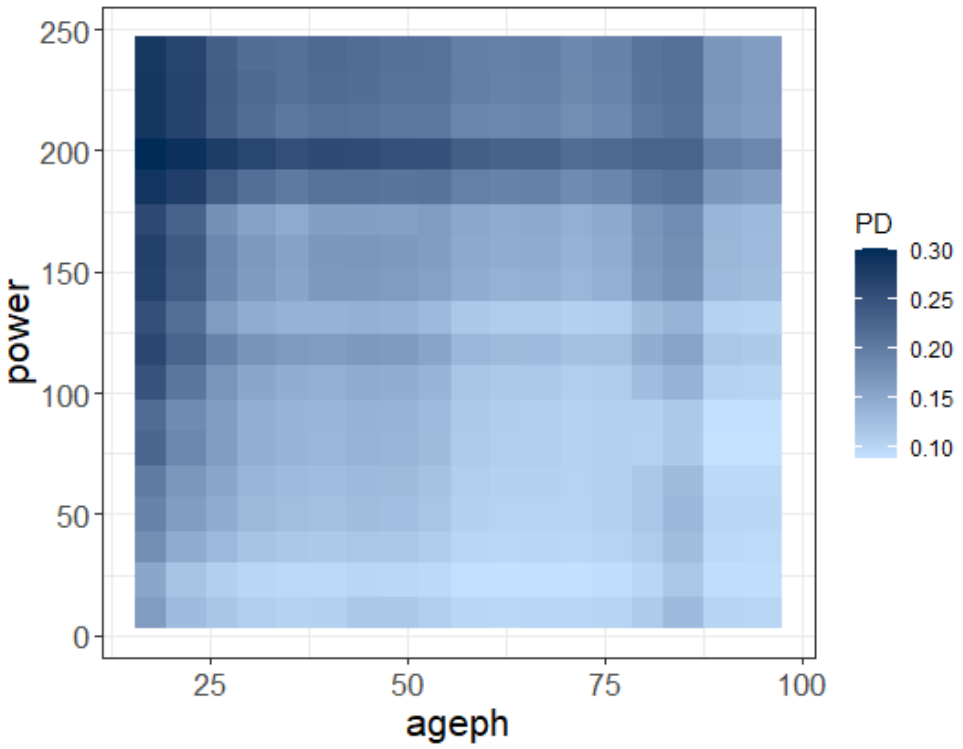




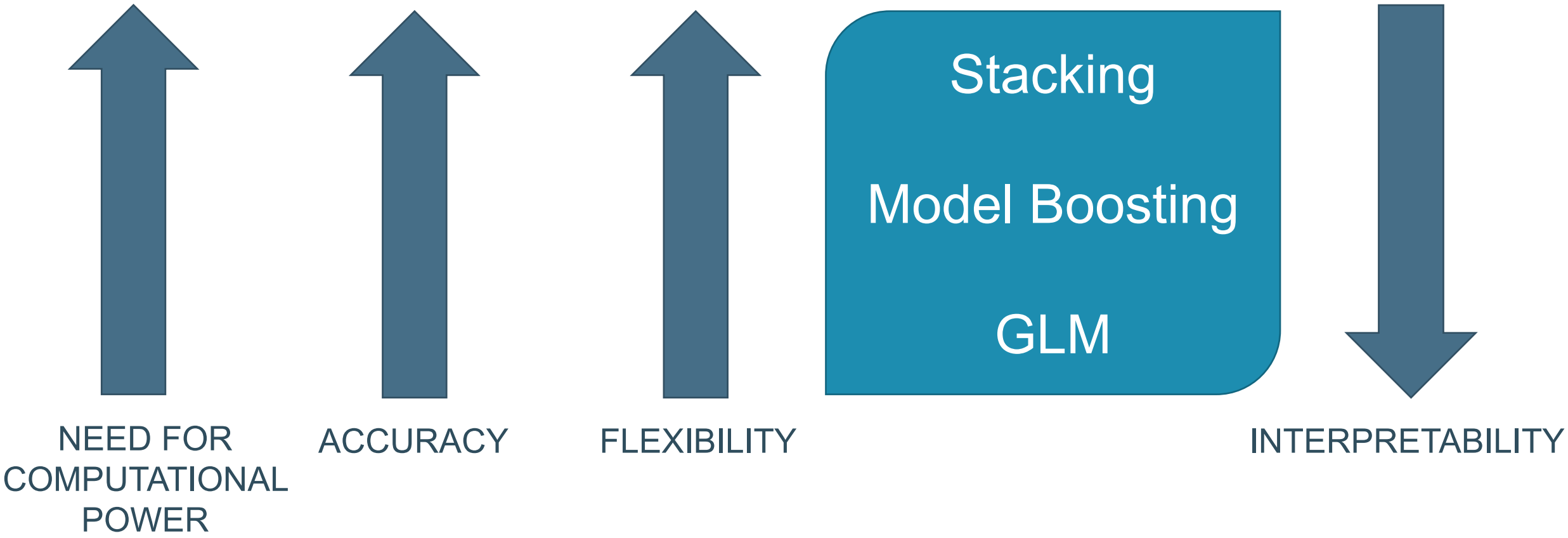


Partial dependence plots





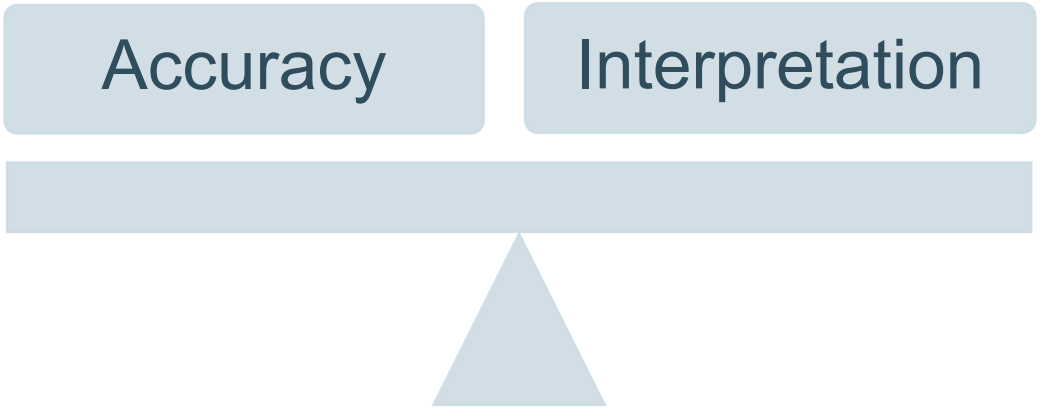
Conclusion


Stacking
Model Boosting
GLM
NEED FOR COMPUTATIONAL POWER
ACCURACY
FLEXIBILITY
INTERPRETABILITY

Future prospects

- Increased data size and complexity
- More powerful machines

→ More complex models



KU LEUVEN

Thank you for your attention!