2nd place solution for Actuarial Loss Prediction

A. Gulyás & N. Fornasin

Team Boosted Goose

Preprocessing

The preprocessing consisted of the following major steps: (purely technical steps are not listed)

- Adjusted unrealistic values of the predictors (e.g. 200 hours worked per week, reporting date before accident date, etc.)
- Added features, such as: weekday of accident, core working hours, reporting delay, etc.
- Excluded observations with inplausible set of predictors

Text analysis

Try to classify sentences based on word occurrence. Weight clusters of words based on median ultimate.

SCRAPER SLIPPED AND HIT HEAD HYPERFLEXION INJURY TO NECK AND SHOULDER

18 13 20 22 23

SLIP	HIT	LEG	HEAD	NECK	KNIFE	SHOULDER	Weight
1	1	0	1	1	0	1	95

Model

The algorithm relied on the following ensemble techniques:

- boosting: gradient boosting using xgboost
- bagging: radom forest as base learner
- voting: custom combination based on insight

Further details:

- natural logarithm as link function
- tweedie distribution of errors
- monotonic constraints for selected features, e.g. WeeklyWages

What worked and what didn't

What worked

- Single word analysis;
- Regression to distribution;
- *Stacking with expert judgement (cooking).*

What didn't work

- Neural networks;
- External data sources (e.g. for inflation);
- *Something about NLP? Like with entity analysis?*