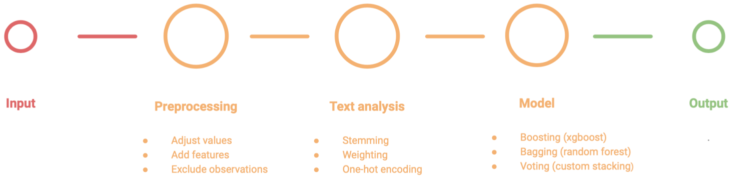
2nd place solution for Actuarial Loss Prediction

A. Gulyás & N. Fornasin

Team Boosted Goose

Model overview

Our model can be divided into three mutually independent blocks:



Main features:

- **Single word analysis**
- **Gradient boosting with random forest**
- **Expert judgement stacking**

Preprocessing

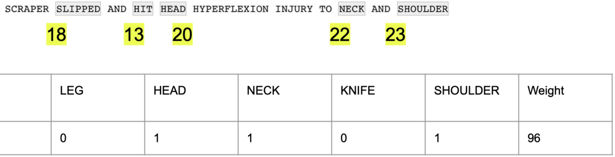
The preprocessing consisted of the following major steps (purely technical steps are not listed):

- **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 implausible set of predictors

Text analysis

Our analysis of the claim description feature:

- **Extraction and stemming** of the most common words (in this step laceration and lacerated both become ”lacer”)
- **Clustering and weighting** of the words according to median ultimate claim cost
- **One hot encoding** for every single word identified



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	96

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