An xgboost solution for Actuarial Loss Prediction

A. Gulyás & N. Fornasin

Team Boosted Goose

Preprocessing

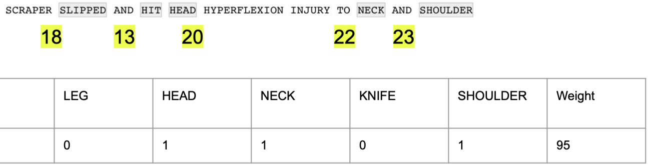
We used pandas because sklearn's pipelines have been designed with the intention of making me angry (it worked). What we did in preprocessing

- Corrected mistakes, such as: 200 hours worked per week, reporting date before accident date...
- Added features, such as: weekday of accident, core working hours, numeric transformations.

It wasn't fancy but it did what it had to, which is more than you can ask.

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

ML Algorithm

After several attempts we decided to focus on a gradient boosted tree. *Write something about ensemble techniques and the number of parallel trees and all these things.*

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?*