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:
- Added features, such as:

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

Text analysis

Of course you can be fancy, but you can also be stubborn, and guess what, stubborn worked

ML Algorithm

Sometimes you can see the forest but not the trees, sometimes the trees but not the forest, here it's been four months and I wasn't able to see either

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?