Proposal of Modeling Approach

Aviation Safety Team 1

Goal 1: Predict injury proportions in each of four categories (none / minor / serious / fatal).

Goal 2: Predict aircraft damage category (minor / serious / destroyed)

* Similar approach for both goals using regression for goal 1 and categorization for goal 2 (e.g. random forest classifier vs. regressor)

General modeling approach

* Start with a random forest for feature selection
  + Try some different hyperparameters, but keep it simple enough that the models will run quickly. The goal here is to quickly eliminate many variables, which should pop out even in simpler models, not to do any fine tuning.
* There are some features that we’ve taken a light approach to cleaning and could be more careful about (e.g. combining different spellings of same make, ‘CESSNA’ = ‘Cessna’ etc.) If any of these seem significant, return to cleaning them more carefully.
* Try a few different models and for each tune hyperparameters with cross-validation
  + Random forest
  + Extra trees
  + XGBoost
  + Bagged k-nearest-neighbors
    - Maybe increase the weight on latitude / longitude in computing distance
* Go with whatever model seems like the best combination of accuracy, speed, interpretability.