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)

Goal 3: Predict the number of accidents in the future using time-series prediction like ARIMA, SARIMA or LSTM

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.
    - Select some number of most “important” variables from feature importances netted from RF classifier
* 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.
  + Jake note: I do wonder if we should do most of this kind of thing beforehand – wouldn’t it be possible that coding errors could mask important relationships?
  + Jinting note: maybe we can try dimension reduction methods like PCA or t-SNE so that we don’t have to decide on the features to keep.
* Try a few different models and for each tune hyperparameters with cross-validation
  + Random forest
  + Extra trees
  + Histogram Gradient Boost Classifier
  + 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.
* Iteratively tune hyperparameters of “best” performing models – whichever combination of hyperparameters performs best, finally deploy to test set