Session 1 Notes

Initial thoughts:

missing values:

* imputation
* negatives as a factor -- is it meaningful?

url parsing:

* length of url - [caleb] looks useful
* url entropy score - [caleb] looks useful
* has 3rd level domain? - [caleb] looks useful
* non alpha characters - [sara]
* non alphanumeric characters - [sara] see if there's a difference but pick one
* top level domain categories - [caleb] inconclusive?

Shape

Description automatically generated with medium confidence

Notes from regroup:

URL features

* made 4 that seem useful:
  + total length of url
  + does it have a third level domain (ie "[www.](http://www.)")
  + url entropy - measures info content of the url which indirectly measures randomness
  + number of non-alphanumeric characters
    - still in progress

imputations

* tried some different approaches (impute with means, drop negatives)
* next step is to try multiple runs of models and see what works best

models to try:

* logistic regression
* naive bayes
* classification tree
* random forest
* KNN
* SVM
* adaboost
* if time: EBM

techniques to try for hyperpameter tuning:

* QDA
* LDA

next steps:

* push code to github
* plan to start trying out models on monday