Homework 3: Common Mistakes

Highest priority mistakes:

- Discretizing or imputing or converting to dummies on the entire data set instead of just doing it on training data
- Train and test sets do not leave gap (as long as the prediction horizon - 60 days for example) in between those sets periods to account for outcome to happen.
- Score thresholds are used to generate precision recall numbers and compare models instead of comparing them on % of projects that are classified as 1 (above a threshold).

Medium priority:

- Throwing away columns that could be useful
- Using all models - High: DT, LR, RF, Medium: Bagging, Boosting (adaboost, Gradient boosting), Extras Trees Low: SVM, kNN, NB
- Using meaningful parameters for each model and varying them

Very minor mistakes:

- Label is 1 for projects that got funded within 60 days instead of those that did not get funded

Repo and Coding:

- Repo Readme
- Modularity - each model is a separate function. Each parameter is a separate loop within that function
- Hard coded/Not reusable
- Repeated code
- Not commented

Rare mistakes:

- Selectbestk has to be done on a holdout set
- Gridsearch cannot use CV and should not optimize for accuracy

Insignificant Parameters:

- SVM is similar to Linear Logistic
- Disadvantages of Automation
    - Selectbestk: enforces double overfitting, as itself finds the best parameters so when we implement other code

- GridsearchCv: does not do consider temporal, you should do that yourself
- According to Rayid: there is no direct relation of number of trees and shrinkage