to_do

9 octobre 2018

To Do to improve model

- walk score & property tranfer & (to lesser extent) pedestrian survey data sets have data points outside city boundaries - this results in corrupted data values along city boundary during grid cell assignment
  - eliminate data points outside city boundaries beifre grid assignment or change grid assignment to only recognize data points inside city limits
- imputation - currently imputed values by median of entire column - improve this to just impute based on values in local geopgrphic area
- boolean variables - not currently inlcuding and boolean / categorical variables - identify variables (like neighborhood) and one-hot encode into data columns
- text mine data values from existing columns – probably some good opportunity here
- property xfer data set - includes information about proerty type (commercial, residential, etc) - include as categorical and one-hot encode
- do a VIF or some other analysis to see if there are columns highly correlated that want to be considered for removal
- are there outliers that need to be removed ?
- add the add/subtract/mulitply/divide feature combinations for each pair of numerically valued columns